THE MULTIPLICITY PROBLEMS FOR THE UNITARY GINZBURG-RALLIS MODELS

CHEN WAN AND LEI ZHANG

Abstract. We consider the local multiplicity problems for the analog of the Ginzburg-Rallis model for unitary groups and unitary similitude groups. For the unitary similitude group case, by proving a local trace formula for the model, we prove a multiplicity formula for all tempered representations, which implies that the summation of the multiplicities is equal to 1 over every tempered local Vogan L-packet. For the unitary group case, we also prove a multiplicity formula for all tempered representations which implies that the summation of the multiplicities is equal to 2 over every tempered local Vogan L-packet.

1. Introduction and Main Results

1.1. **Main results.** Let F be a nonarchimedean field of characteristic 0 and $E = F(\sqrt{\alpha})$ be a quadratic extension of F. Let $\eta_{E/F}\colon F^\times \to \mathbb{C}^\times$ be the quadratic character associated to E via the local class field theory, $N_{E/F}$ (resp. $\mathrm{tr}_{E/F}$) be the norm map (resp. trace map), and $x \to \bar{x}$ be the Galois action on E. Denote w_n to be the symmetric matrix of size $n \times n$ given by

$$w_n = \begin{pmatrix} & w_{n-1} \\ 1 & \end{pmatrix} \text{ and } w_1 = (1).$$

For $\varepsilon \in F^\times$, let

$$J_{2n,\varepsilon} = \begin{pmatrix} 0 & 0 & w_{n-1} \\ 0 & A_\varepsilon & 0 \\ w_{n-1} & 0 & 0 \end{pmatrix} \text{ where } A_\varepsilon = \begin{pmatrix} -\varepsilon & 0 \\ 0 & 1 \end{pmatrix}.$$

Define the unitary similitude group $\mathrm{GU}_{2n,\varepsilon}(F) = \mathrm{GU}(J_{2n,\varepsilon})(F)$ to be

$$\mathrm{GU}(J_{2n,\varepsilon})(F) = \{g \in \mathrm{GL}_{2n}(E)\colon {}^t\bar{g}J_{2n,\varepsilon}g = \lambda(g)J_{2n,\varepsilon}\} \tag{1.1}$$

where $\lambda(g) \in F^\times$ is the similitude factor of g. Note that if $\varepsilon \in \mathrm{Im}(N_{E/F})$, then $\mathrm{GU}(J_{2n,\varepsilon})$ is quasi-split; if $\varepsilon \notin \mathrm{Im}(N_{E/F})$, then $\mathrm{GU}(J_{2n,\varepsilon})$ is the non-quasi-split inner form of the quasi-split unitary similitude group. In this paper, we mainly work on the groups $G_\varepsilon = \mathrm{GU}(J_{6,\varepsilon})$.

2010 *Mathematics Subject Classification.* Primary 22E35, 22E50.

Key words and phrases. Harmonic Analysis on Spherical Variety, Representation of p-adic Group, Local Trace Formula, Multiplicity One on Vogan Packet.

Next, we introduce a spherical subgroup of G_ε. Let $P_\varepsilon = M_\varepsilon U_\varepsilon$ be the standard parabolic subgroup of $\mathrm{GU}(J_{6,\varepsilon})$ with

$$M_\varepsilon(F) = \{m(g,h) = \begin{pmatrix} g & & \\ & h & \\ & & \lambda(h)g^* \end{pmatrix} : g \in \mathrm{GL}_2(E),\ g^* = w_2{}^t\bar{g}^{-1}w_2,\ h \in \mathrm{GU}(J_{2,\varepsilon})(F)\},$$

$$U_\varepsilon(F) = \{u(X,Y) = \begin{pmatrix} I_2 & X & Y \\ & I_2 & X' \\ & & I_2 \end{pmatrix} : X, Y \in Mat_{2\times 2}(E), X' = -A_\varepsilon^{-1}{}^tXw_2,$$
$$w_2Y + {}^t\bar{Y}w_2 + {}^tX'A_\varepsilon X' = 0\}.$$

Let ξ_ε be a generic character of $U_\varepsilon(F)$ given by

$$\xi_\varepsilon(u(X,Y)) = \psi(\mathrm{tr}_{E/F}(\mathrm{tr}(X)))$$

where ψ is a non-trivial additive character of F. Then the stabilizer of ξ_ε under the adjoint action of $M_\varepsilon(F)$ is

$$H_{0,\varepsilon}(F) := \{m(h,h) \colon h \in \mathrm{GU}(J_{2,\varepsilon})(F)\}.$$

Let χ_F (resp. χ_E) be a character of $F^\times$ (resp. $E^\times$). We then define the character ω_ε of $H_{0,\varepsilon}(F)$ to be

$$\omega_\varepsilon(m(h,h)) = \chi_E(\det(h))\chi_F(\lambda(h))$$

where λ is the similitude character of $\mathrm{GU}(J_{2,\varepsilon})(F)$. Let η be the restriction of the character ω_ε to the center $Z_{H_{0,\varepsilon}}(F) = Z_{G_\varepsilon}(F) \simeq E^\times$. It is easy to see that $\eta = \chi_E^2 \otimes (\chi_F \circ N_{E/F})$.

Define $H_\varepsilon = H_{0,\varepsilon} \ltimes U_\varepsilon$, which is a spherical subgroup of $\mathrm{GU}(J_{6,\varepsilon})$. Then we have a character $\omega_\varepsilon \otimes \xi_\varepsilon$ of $H_\varepsilon(F)$. Let π_ε be a smooth finite length representation of $G_\varepsilon(F)$ with central character η. We define the multiplicity

$$m(\pi_\varepsilon) = \dim(\mathrm{Hom}_{H_\varepsilon(F)}(\pi_\varepsilon, \omega_\varepsilon \otimes \xi_\varepsilon)).$$

The goal of this paper is to study the behavior of the multiplicity $m(\pi_\varepsilon)$ over the local Vogan L-packet.

For $i = 1, 2$, fix $\varepsilon_i \in F^\times$ with $\eta_{E/F}(\varepsilon_i) = (-1)^{i-1}$. Let ϕ be a tempered Langlands parameter for $\mathrm{GU}_6(F)$. Assume the endoscopic classification holds for even unitary similitude groups (This is expected from the endoscopic classification of unitary groups in [10] and [9], together with Xu's work [24] on the reduction from the similitude classical groups to classical groups. We refer the reader to Section 2.6 for details). Then the parameter ϕ determines a tempered local Vogan L-packet $\Pi_\phi = \Pi_\phi(G_{\varepsilon_1}) \cup \Pi_\phi(G_{\varepsilon_2})$ consisting of a finite number of tempered representations of $G_{\varepsilon_1}(F)$ and $G_{\varepsilon_2}(F)$ respectively. Our main theorem can be stated as follows.

Theorem 1.1. *For all tempered Langlands parameters ϕ of $\mathrm{GU}_6(F)$ with central character η, we have*

$$\sum_{i=1}^{2} \sum_{\pi_{\varepsilon_i} \in \Pi_\phi(G_{\varepsilon_i})} m(\pi_{\varepsilon_i}) = 1.$$

In other words, the summation of the multiplicities over every tempered local Vogan L-packet is equal to 1.

Then we study the analog of the model $(G_\varepsilon, H_\varepsilon)$ for unitary groups. For $\varepsilon \in F^\times$, we define the unitary group $\mathrm{U}(J_{2n,\varepsilon})$ to be

$$\mathrm{U}(J_{2n,\varepsilon})(F) = \{g \in \mathrm{GL}_{2n}(E) \colon {}^t\bar{g} J_{2n,\varepsilon} g = J_{2n,\varepsilon}\}. \tag{1.2}$$

Let $G_{1,\varepsilon} = \mathrm{U}(J_{6,\varepsilon})$. As in the similitude case, we can define the subgroups $H_{1,\varepsilon} = H_{0,1,\varepsilon} \ltimes U_{1,\varepsilon}$ of $G_{1,\varepsilon}$ with $H_{0,1,\varepsilon}(F) \simeq \mathrm{U}(J_{2,\varepsilon})(F)$ and $U_{1,\varepsilon} = U_\varepsilon$. We can also define the character $\omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon}$ of $H_{1,\varepsilon}(F)$ via the characters ψ and χ_E (note that here we don't have similitude character, hence we can only define the character $\omega_{1,\varepsilon}$ via the determinant map). Let η_1 be the restriction of the character $\omega_{1,\varepsilon}$ to the center $Z_{H_{0,1,\varepsilon}}(F) = Z_{G_{1,\varepsilon}}(F) \simeq E^1$ where E^1 is the kernel of the norm map $N_{E/F}$. It is easy to see that $\eta_1 = \chi_E^2|_{E^1}$. Let $\pi_{1,\varepsilon}$ be a smooth finite length representation of $G_{1,\varepsilon}(F)$ with central character η_1, we define the multiplicity

$$m(\pi_{1,\varepsilon}) = \dim(\mathrm{Hom}_{H_{1,\varepsilon}(F)}(\pi_{1,\varepsilon}, \omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon})).$$

For $i = 1, 2$, let $\varepsilon_i \in F^\times$ with $\eta_{E/F}(\varepsilon_i) = (-1)^{i-1}$ as before. Let ϕ be a tempered Langlands parameter for $\mathrm{U}_6(F)$. By the endoscopic classification of unitary groups in [10] and [9], the parameter ϕ determines a tempered local Vogan L-packet $\Pi_\phi = \Pi_\phi(G_{1,\varepsilon_1}) \cup \Pi_\phi(G_{1,\varepsilon_2})$ consisting of a finite number of tempered representations of $G_{1,\varepsilon_1}(F)$ and $G_{1,\varepsilon_2}(F)$ respectively. Our main theorem for the unitary group case can be stated as follows.

Theorem 1.2. *For all tempered Langlands parameters ϕ of $\mathrm{U}_6(F)$ with central character η_1, we have*

$$\sum_{i=1}^{2} \sum_{\pi_{1,\varepsilon_i} \in \Pi_\phi(G_{1,\varepsilon_i})} m(\pi_{1,\varepsilon_i}) = 2.$$

In other words, the summation of the multiplicities over every tempered local Vogan L-packet is equal to 2.

Remark 1.3. *The models $(G_\varepsilon, H_\varepsilon)$ (resp. $(G_{1,\varepsilon}, H_{1,\varepsilon})$) can be viewed as the analog of the Ginzburg–Rallis model (GR for simplicity) for unitary similitude groups (resp. unitary groups). The local multiplicity problem for the Ginzburg–Rallis model has been considered by the first named author in* [17], [18], [19] *and* [21]. *We refer the reader to* [20] *for the definition of the model and the results.*

Remark 1.4. *We expect the results in Theorems 1.1 and 1.2 hold for all generic local Vogan L-packets. For the unitary similitude group case, we also expect the model $(G_\varepsilon, H_\varepsilon)$ to be a Gelfand pair, i.e. $m(\pi_\varepsilon) \leq 1$ for all irreducible smooth representations of $G_\varepsilon(F)$. Theorem 1.1 verifies this inequality for all tempered representations. However, the model $(G_{1,\varepsilon}, H_{1,\varepsilon})$ is not a Gelfand pair. In fact, later in our proof, we can show that when $G_{1,\varepsilon}$ is quasi-split, $m(\pi_{1,\varepsilon}) = 2$ for all unramified tempered principal series* *of $G_{1,\varepsilon}(F)$. For this model, we expect that the multiplicity is always less or equal to* 2*.*

Remark 1.5. *As in the Ginzburg–Rallis model case, globally we expect the period integrals of the models $(G_\varepsilon, H_\varepsilon)$ to be related to the central value of the exterior cube L-function of* GU_6*. We refer the reader to Section 1 of* [23] *for details.*

1.2. **Remarks on the proofs.** We first discuss the proof of Theorem 1.1 (i.e. the unitary similitude group case). Our proof of Theorem 1.1 uses the method developed by Waldspurger in his proof of the local orthogonal Gan–Gross–Prasad (GGP for simplicity) conjecture ([15], [16]). To be specific, we are going to prove a multiplicity formula

$$m(\pi_\varepsilon) = m_{geom}(\pi_\varepsilon)$$

for all the tempered representations π_ε of $G_\varepsilon(F)$. Here $m_{geom}(\pi_\varepsilon)$ is defined in terms of the regular germs of the distribution character θ_{π_ε} of π_ε. Theorem 1.1 will follow from the multiplicity formula together with the behavior of the distribution characters on the local L-packet. It is worth to mention that Waldspurger's method later has been adapted by Beuzart-Plessis for the unitary GGP model ([1], [2]), and by the first named author for the Ginzburg–Rallis model ([17], [18]). It has also been used in [3] and [5] for the Galois model and the generalized Shalika model (but with different proofs of the geometric side of the trace formula).

In order to prove the multiplicity formula, as in all the previous cases, one needs to prove a local trace formula. We refer the reader to Section 4 for the statements of the trace formula and the multiplicity formula. Our proof for the geometric side of the trace formula is quite similar to the GGP case in [2] and all the computations are very similar to the GR case in [17]. As a result, we will only give a sketch of the proof without providing details (see Section 4.3).

As for the spectral side of the trace formula, our proof is quite different from the GGP case and the GR case. The main reason is that unlike the previous cases, we don't have the Gelfand pair condition for the model $(G_\varepsilon, H_\varepsilon)$ (although it is expected, see Remark 1.4). To avoid using the Gelfand pair condition, we decompose the Harish-Chandra–Schwartz space $\mathcal{C}(G(F))$ into two subspaces $\mathcal{C}(G(F)) = {}^\circ\mathcal{C}(G(F)) \oplus \mathcal{C}_{ind}(G(F))$ where ${}^\circ\mathcal{C}(G(F))$ corresponds to the discrete series and $\mathcal{C}_{ind}(G(F))$ corresponds to the induced representations. We prove the spectral expansion for these two subspaces separately. For the subspace ${}^\circ\mathcal{C}(G(F))$, we uses the method developed by Beuzart-Plessis for the Galois model case in [3] which does not require the Gelfand pair condition. For the space $\mathcal{C}_{ind}(G(F))$, we first prove the multiplicity one result for all the reduced models. Then in Appendix A, by applying Mackey theory, we can show that the Gelfand pair condition holds for all tempered representations that are not discrete series (Proposition 4.14). This allows us to prove the spectral expansion for the subspace $\mathcal{C}_{ind}(G(F))$

by applying the same argument as in the GGP case. For details, see Section 6.

Now we move to the proof of Theorem 1.2 (i.e. the unitary group case). The idea is still to prove a multiplicity formula for all the tempered representations and then prove the theorem by applying the multiplicity formula together with the behavior of the distribution characters on the local L-packet. However, the proof of the multiplicity formula is quite different from all the previous cases. To be specific, in all the previous cases, the proof of the multiplicity formula is based on the proof of a local trace formula for the model. However, for the model $(G_{1,\varepsilon}, H_{1,\varepsilon})$, because of the following two issues, it is not clear to us how to prove the local trace formula.

(1) There are more than one open Borel orbit (in fact, there are two of them) for the model $(G_{1,\varepsilon}(F), H_{1,\varepsilon}(F))$.
(2) For the slice representation (i.e. the conjugation action of $H_{1,\varepsilon}(F)$ on the tangent space of the spherical variety $G_{1,\varepsilon}(F)/H_{1,\varepsilon}(F)$ at the identity element 1), the regular orbits do not correspond to the orbits under the $G_{1,\varepsilon}(F)$-conjugation. Some $G_{1,\varepsilon}(F)$ conjugacy class in the tangent space will break into two $H_{1,\varepsilon}(F)$ conjugacy classes.

The first issue creates some difficulties for the spectral side, while the second issue creates difficulties for the geometric side.

Remark 1.6. *By the discussion in Section 5 of* [22]*, we know that both issues above are due to the fact that the number* $d(G_{1,\varepsilon}, H_{1,\varepsilon}, F)$ *(defined in Definition 5.5 of* [22]*) for the model* $(G_{1,\varepsilon}, H_{1,\varepsilon})$ *is equal to 2 (Lemma 7.8 of* [22]*). For all the other known cases, this number is equal to 1. As far as we know, this is the first case (and the only case so far) where the multiplicity formula has been proved when this number is greater than 1.*

As a result, we prove the multiplicity formula by a different method. To be specific, we first prove a relation between the model $(G_{1,\varepsilon}, H_{1,\varepsilon})$ and the model $(G_{\varepsilon}, H_{\varepsilon})$ (see Proposition 5.2). Then we prove the multiplicity formula of the model $(G_{1,\varepsilon}, H_{1,\varepsilon})$ by applying the multiplicity formula of the model $(G_{\varepsilon}, H_{\varepsilon})$ together with Proposition 5.2. For details, see Section 5.2.

The last thing we want to emphasize about the unitary group case is that in the multiplicity formula for the unitary group case, the regular germ at the identity element has coefficient 2 and this is why we have the summation of the multiplicities over the L-packet is equal to 2. This is different from all the previous known cases where the regular germs have coefficient 1. This new phenomenon is again due to the fact that the number $d(G_{1,\varepsilon}, H_{1,\varepsilon}, F)$ (defined in Definition 5.5 of [22]) for the model $(G_{1,\varepsilon}, H_{1,\varepsilon})$ is equal to 2.

1.3. **Organizations of the paper.** In Section 2, we introduce basic notation and conventions of this paper. We will also discuss some local representation theory of the unitary groups and the unitary similitude groups. In Section 3, we study the analytic and geometric properties of the model

$(G_\varepsilon, H_\varepsilon)$. Then we discuss some estimates for various integrals which will be used in later sections.

In Section 4, we state the trace formula and multiplicity formula for the model $(G_\varepsilon, H_\varepsilon)$ and for all its reduced models. In Section 4.3, we give a sketch of the proof of the geometric side of the trace formula. Since the idea of the proof is similar to the GGP case and all the computations are very similar to the GR case, we will skip the details of the proof. Then in Section 4.5 and 4.6, we consider the trace formula and multiplicity formula for reduced models. We will postpone the proof of a technical proposition (i.e. Proposition 4.14) to Appendix A.

In Section 5, we prove our main theorems by assuming the trace formula holds. Then in Section 6, we prove the trace formula. Finally, in Appendix A, we prove Proposition 4.14 by using Mackey theory.

1.4. **Acknowledgement.** We would like to thank Dihua Jiang for suggesting us thinking about this problem. Collaboration on this work started at a workshop in the American Institute of Mathematics, and the work was completed at a workshop in Zhejiang University. We thank these institutions for their hospitality and the organizers for the invitations. We would also like to thank an anonymous referee for many helpful comments and corrections. The work of the first author is partially supported by the NSF grant DMS-2103720. The work of the second author is partially supported by AcRF Tier 1 grants R-146-000-237-114, R-146-000-277-114, and R-146-000-308-114 of National University of Singapore.

2. Preliminaries

2.1. **Notation and conventions.** Let F be a p-adic field, and let $|\cdot| = |\cdot|_F$ be the absolute value on F. For every connected reductive group G defined over F, let Z_G be the center of G and A_G be the maximal split torus of Z_G. We denote by $X(G)$ the group of F-rational characters of G. Define $\mathfrak{a}_G = \mathrm{Hom}(X(G), \mathbb{R})$, and let $\mathfrak{a}_G^* = X(G) \otimes_{\mathbb{Z}} \mathbb{R}$ be the dual of $\mathfrak{a}_G$. We define a homomorphism $H_G : G(F) \to \mathfrak{a}_G$ by $H_G(g)(\chi) = \log(|\chi(g)|_F)$ for $g \in G(F)$ and $\chi \in X(G)$. Let $\mathfrak{a}_{G,F}$ (resp. $\tilde{\mathfrak{a}}_{G,F}$) be the image of $G(F)$ (resp. $Z_G(F)$) under H_G. Then $\mathfrak{a}_{G,F}$ and $\tilde{\mathfrak{a}}_{G,F}$ are lattices in $\mathfrak{a}_G$. Let $\mathfrak{a}_{G,F}^\vee = \mathrm{Hom}(\mathfrak{a}_{G,F}, 2\pi\mathbb{Z})$ and let $\tilde{\mathfrak{a}}_{G,F}^\vee = \mathrm{Hom}(\tilde{\mathfrak{a}}_{G,F}, 2\pi\mathbb{Z})$. Set $\mathfrak{a}_{G,F}^* = \mathfrak{a}_G^*/\mathfrak{a}_{G,F}^\vee$. We can identify $i\mathfrak{a}_{G,F}^*$ with the group of unitary unramified characters of $G(F)$ by letting $\lambda(g) = e^{\langle \lambda, H_G(g)\rangle}$ for $\lambda \in i\mathfrak{a}_{G,F}^*$ and $g \in G(F)$. For a Levi subgroup M of G, let $\mathfrak{a}_{M,0}^*$ be the subset of elements in $\mathfrak{a}_{M,F}^*$ whose restriction to $\tilde{\mathfrak{a}}_{G,F}$ is zero. Then we can identify $i\mathfrak{a}_{M,0}^*$ with the group of unitary unramified characters of $M(F)$ which are trivial on $Z_G(F)$.

Let $\mathfrak{g}$ be the Lie algebra of G. For a Levi subgroup M of G, let $\mathcal{P}(M)$ be the set of parabolic subgroups of G whose Levi part is M, $\mathcal{L}(M)$ be the set of Levi subgroups of G containing M, and let $\mathcal{F}(M)$ be the set of parabolic subgroups of G containing M. We have a natural decomposition

$\mathfrak{a}_M = \mathfrak{a}_M^G \oplus \mathfrak{a}_G$. Denote by $proj_M^G$ and $proj_G$ the projections of $\mathfrak{a}_M$ to each factors. Fix a special maximal compact subgroup K of G. We have the Iwasawa decomposition $G(F) = M(F)U(F)K$. For each $P \in \mathcal{P}(M)$, we can associate a positive chamber $\mathfrak{a}_P^+ \subset \mathfrak{a}_M$, and we can also define a function $H_P : G(F) \to \mathfrak{a}_M$ by $H_P(g) = H_M(m_g)$ where $g = m_g u_g k_g$ is the Iwasawa decomposition of g.

Let $\|\cdot\|$ be a height function on $G(F)$, taking values in $\mathbb{R}_{\geq 1}$. We say two height functions $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent if there exists $C, N > 0$ such that

$$C^{-1}(\|g\|_1)^N \leq \|g\|_2 \leq C(\|g\|_1)^N,\ \forall g \in G(F).$$

If $G = \mathrm{GL}_n$, we can define the height function as

$$\|g\| = \max\{|x_{ij}|,\ |y_{ij}|,\ 1 \leq i,j \leq n\},\ g = (x_{ij})_{1\leq i,j\leq n}, g^{-1} = (y_{ij})_{1\leq i,j\leq n}.$$

In general, we have a closed embedding from G into GL_n and this induces a height function on G. The height function is independent of the choice of the embedding in the sense that the height functions defined by two different embeddings are equivalent to each other.

We define a log-norm σ on $G(F)$ by $\sigma(g) = \sup\{1, \log(\|g\|)\}$. We also define $\sigma_0(g) = \inf_{z\in Z_G(F)}\{\sigma(zg)\}$. Similarly, we can define the log-norm function on $\mathfrak{g}(F)$: fixing a basis $\{X_i\}$ of $\mathfrak{g}(F)$, for $X \in \mathfrak{g}(F)$, let $\sigma(X) = \sup\{1, \sup\{\log(|a_i|)\}\}$, where a_i are the X_i-coordinate of X.

Let M_{min} be a minimal parabolic subgroup of G and $A_{min} = A_{M_{min}}$. For each $P_{min} \in \mathcal{P}(M_{min})$, let $\Psi(A_{min}, P_{min})$ be the set of positive roots associated to P_{min}, and let $\Delta(A_{min}, P_{min}) \subset \Psi(A_{min}, P_{min})$ be the subset of simple roots.

For $x \in G$ (resp. $X \in \mathfrak{g}$), let $Z_G(x)$ (resp. $G_X = Z_G(X)$) be the centralizer of x (resp. X) in G, and let G_x be the neutral component of $Z_G(x)$. Accordingly, let $\mathfrak{g}_x$ (resp. $\mathfrak{g}_X$) be the Lie algebra of G_x (resp. G_X). Denote by $G_{ss}(F)$ the set of semisimple elements in $G(F)$, and by $G_{reg}(F)$ the set of regular semisimple elements in $G(F)$. The Lie algebra versions are denoted by $\mathfrak{g}_{ss}(F)$ and $\mathfrak{g}_{reg}(F)$, respectively. For $x \in G_{ss}(F)$ (resp. $X \in \mathfrak{g}_{ss}(F)$), let $D^G(x)$ (resp. $D^G(X)$) be the Weyl determinant. We say $x \in G_{ss}(F)$ is elliptic if $A_G = A_{G_x}$, i.e. the maximal split torus of Z_{G_x} is equal to the maximal split torus of Z_G. We use $G(F)_{ell}$ to denote the set of elliptic elements in $G(F)$ and we use $G(F)_{ell,reg} = G(F)_{ell} \cap G_{reg}(F)$ to denote the set of regular elliptic elements in $G(F)$. Similarly we can also define $\mathfrak{g}(F)_{ell}$ and $\mathfrak{g}(F)_{ell,reg}$.

For two complex valued functions f and g on a set X with g taking values in the positive real numbers, we write $f(x) \ll g(x)$, and say that f is *essentially bounded* by g, if there exists a constant $c > 0$ such that for all $x \in X$, we have $|f(x)| \leq cg(x)$. We say f and g are *equivalent*, which is denoted by $f(x) \sim g(x)$, if f is essentially bounded by g and g is essentially bounded by f.

Lastly, we fix a non-trivial additive character $\psi : F \to \mathbb{C}^\times$. If G is a connected reductive group, we may fix a non-degenerate symmetric bilinear form $\langle\cdot,\cdot\rangle$ on $\mathfrak{g}(F)$ that is invariant under $G(F)$-conjugation (i.e. Killing form). For $f \in C_c^\infty(\mathfrak{g}(F))$, we can define its Fourier transform $f \to \hat{f}$ to be

$$\hat{f}(X) = \int_{\mathfrak{g}(F)} f(Y)\psi(\langle X, Y\rangle)\,\mathrm{d}Y \tag{2.1}$$

where $\mathrm{d}Y$ is the self-dual Haar measure on $\mathfrak{g}(F)$ such that $\hat{\hat{f}}(X) = f(-X)$. Then we get a Haar measure on $G(F)$ such that the Jacobian of the exponential map equals 1. If H is a subgroup of G such that the restriction of the bilinear form to $\mathfrak{h}(F)$ is also non-degenerate, then we can define the measures on $\mathfrak{h}(F)$ and $H(F)$ by the same method.

2.2. **Induced representation.** Given a parabolic subgroup $P = MU$ of G and a smooth finite length representation (τ, V_τ) of $M(F)$, let $(I_P^G(\tau), I_P^G(V_\tau))$ be the normalized parabolic induced representation: $I_P^G(V_\tau)$ is the space of smooth functions $e\colon G(F) \to V_\tau$ such that

$$e(mug) = \delta_P(m)^{1/2}\tau(m)e(g),\ m \in M(F),\ u \in U(F),\ g \in G(F).$$

And the $G(F)$-action is just the right translation. Here $\delta_P : M(F) \to \mathbb{C}^\times$ is the modular character.

For $\lambda \in \mathfrak{a}_M^* \otimes_{\mathbb{R}} \mathbb{C}$, let τ_λ be the unramified twist of τ, i.e. $\tau_\lambda(m) = \exp(\lambda(H_M(m)))\tau(m)$ and let $I_P^G(\tau_\lambda)$ be the induced representation. By the Iwasawa decomposition, every function $e \in I_P^G(\tau_\lambda)$ is determined by its restriction on K (recall that K is a special maximal compact subgroup of $G(F)$), and that space is invariant under the unramified twist. That is, for any λ, we can realize the representation $I_P^G(\tau_\lambda)$ on the space $I_{K\cap P}^K(\tau_K)$ which consists of functions $e_K : K \to V_\tau$ such that (note that δ_P is trivial on $M(F) \cap K$)

$$e(mug) = \tau(m)e(g),\ m \in M(F) \cap K,\ u \in U(F) \cap K,\ g \in K.$$

Here τ_K is the restriction of τ to the group $K \cap M(F)$.

2.3. **The Harish-Chandra–Plancherel formula.** Let Ξ^G be the Harish-Chandra function of G and let $\mathcal{C}(G(F))$ (resp. $\mathcal{C}^w(G(F))$) be the Harish-Chandra–Schwartz space (resp. weak Harish-Chandra–Schwartz space) of G. We refer the reader to Section 1.5 of [2] for the definitions and basic properties of Ξ^G and the (weak) Harish-Chandra–Schwartz space. Given a unitary character χ of $Z_G(F)$, we define the Harish-Chandra–Schwartz space $\mathcal{C}(Z_G(F)\backslash G(F), \chi)$ to be the Mellin transform of the space $\mathcal{C}(G(F))$ with respect to χ, i.e. functions in $\mathcal{C}(Z_G(F)\backslash G(F), \chi)$ are of the form

$$f(g) = \int_{Z_G(F)} f'(gz)\chi^{-1}(z)dz$$

with $f' \in \mathcal{C}(G(F))$. We also define the weak Harish-Chandra–Schwartz space $\mathcal{C}^w(Z_G(F)\backslash G(F), \chi)$ to be

$$\mathcal{C}^w(Z_G(F)\backslash G(F), \chi) = \{f \in \mathcal{C}^w(G(F)) |\ f(zg) = \chi(z)f(g) \text{ for all } g \in G(F), z \in Z_G(F)\}.$$

For $d>0$, let $\mathcal{C}_d^w(G(F))$ be the subspace of $\mathcal{C}^w(G(F))$ defined in Section 1.5 of [2] and we let $\mathcal{C}_d^w(Z_G(F)\backslash G(F),\chi)=\mathcal{C}_d^w(G(F))\cap \mathcal{C}^w(Z_G(F)\backslash G(F),\chi)$.

Fix a unitary character χ of $Z_G(F)$. For every $M\in \mathcal{L}(M_{min})$, fix an element $P\in \mathcal{P}(M)$. Let $\Pi_2(M,\chi)$ be the set of isomorphism classes of irreducible discrete series of $M(F)$ whose central character agrees with χ on $Z_G(F)$. Then $i\mathfrak{a}_{M,0}^{*}$ acts on $\Pi_2(M,\chi)$ by the unramified twist. Let $\{\Pi_2(M,\chi)\}$ be the set of orbits under this action. For every orbit $\mathcal{O}$, and for a fixed $\tau\in\mathcal{O}$, let $i\mathfrak{a}_{\mathcal{O}}^{\vee}$ be the set of $\lambda\in i\mathfrak{a}_{M,0}^{*}$ such that the representations τ and τ_λ are equivalent, which is a finite set. We say $(M,\mathcal{O})$ and $(M',\mathcal{O}')$ are equivalent if there exists $g\in G(F)$ such that the g-conjugation map sends M to M' and $\mathcal{O}$ to $\mathcal{O}'$. We use $\sim$ to denote this equivalent relation.

For $f\in \mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1})$, the Harish-Chandra–Plancherel formula ([14]) is

$$f(g)=\sum_{(M,\mathcal{O})\in\{M\in\mathcal{L}(M_{min}),\mathcal{O}\in\{\Pi_2(M,\chi)\}\}/\sim}|W^M||W^G|^{-1}|i\mathfrak{a}_{\mathcal{O}}^{\vee}|^{-1}\int_{i\mathfrak{a}_{M,0}^{*}}\mu(\tau_\lambda)\mathrm{tr}(I_P^G(\tau_\lambda)(g^{-1})I_P^G(\tau_\lambda)(f))\,\mathrm{d}\lambda.$$

Here W^G (resp. W^M) is the Weyl group of G (resp. M), $d\lambda$ is the Haar measure on $i\mathfrak{a}_{M,0}^{*}$ so that the total volume is equal to 1 (note that $i\mathfrak{a}_{M,0}^{*}$ is compact), and $\mu(\tau_\lambda)d\lambda$ is the Plancherel measure.

To simplify our notation, let

$$\Pi_{temp}(G,\chi)=\{I_P^G(\tau)|\ (M,\mathcal{O})\in\{M\in\mathcal{L}(M_{min}),\mathcal{O}\in\{\Pi_2(M,\chi)\}\}/\sim,\ \tau\in\mathcal{O}\}.$$

We define a Borel measure $\mathrm{d}\pi$ on $\Pi_{temp}(G,\chi)$ such that

$$\int_{\Pi_{temp}(G,\chi)}\varphi(\pi)\,\mathrm{d}\pi=\sum_{(M,\mathcal{O})\in\{M\in\mathcal{L}(M_{min}),\mathcal{O}\in\{\Pi_2(M,\chi)\}\}/\sim}|W^M|\cdot|W^G|^{-1}|i\mathfrak{a}_{\mathcal{O}}^{\vee}|^{-1}\int_{i\mathfrak{a}_{M,0}^{*}}\varphi(I_P^G(\tau_\lambda))\,\mathrm{d}\lambda$$

for every compactly supported function φ on $\Pi_{temp}(G,\chi)$. Here by saying a function φ is compactly supported on $\Pi_{temp}(G,\chi)$ we mean that it is supported on finitely many orbits $\mathcal{O}$. Then the Harish-Chandra–Plancherel formula above becomes

$$f(g)=\int_{\Pi_{temp}(G,\chi)}\mathrm{tr}(\pi(g^{-1})\pi(f))\mu(\pi)\,\mathrm{d}\pi.$$

We also need a stronger version of the Harish-Chandra–Plancherel formula (also called matricial Paley-Wiener theorem). As in Section 2.2 of [2], for $\pi\in\Pi_{temp}(G,\chi)$, we equip π with the finest locally convex topology and let $\mathrm{End}(\pi)$ be the space of continuous endomorphisms of the space of π. It is a continuous representation of $G(F)\times G(F)$ via the left and right translation. Let $\mathrm{End}(\pi)^\infty$ be the subspace of smooth vectors in $\mathrm{End}(\pi)$. Let $C^\infty(\Pi_{temp}(G,\chi))$ be the space of functions $\pi\in\Pi_{temp}(G,\chi)\to T_\pi\in \mathrm{End}(\pi)^\infty$ satisfying the following condition:

- for every $(M,\mathcal{O})\in\{M\in\mathcal{L}(M_{min}),\mathcal{O}\in\{\Pi_2(M,\chi)\}\}/\sim$, the restriction of the function to the component $\{\pi_\lambda=I_P^G(\tau_\lambda)|\ \lambda\in i\mathfrak{a}_{M,0}^{*}\}$

is a smooth function from $i\mathfrak{a}_{M,0}^*$ to a finite dimensional subspace of $\mathrm{End}(\pi_\lambda)^\infty \simeq \mathrm{End}(\pi_K)^\infty$ where $\pi_K = \pi_\lambda|_K$.

We define $\mathcal{C}(\Pi_{temp}(G,\chi))$ to be a subspace of $C^\infty(\Pi_{temp}(G,\chi))$ consisting of those $T : \pi \to T_\pi$ such that T is nonzero on finitely many orbits $\mathcal{O}$. Then the matricial Paley-Wiener Theorem ([14]) states that we have an isomorphism between $\mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1})$ and $\mathcal{C}(\Pi_{temp}(G,\chi))$ given by

$$f \in \mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1}) \to T_f := (\pi \in \Pi_{temp}(G,\chi) \to \pi(f) \in \mathrm{End}(\pi)^\infty),$$

$$T \in \mathcal{C}(\Pi_{temp}(G,\chi)) \to f_T(g) = \int_{\Pi_{temp}(G,\chi)} \mathrm{tr}(\pi(g^{-1})T_\pi)\mu(\pi)\,\mathrm{d}\pi.$$

Finally, we introduce two subspaces of $\mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1})$. Let $\Pi_2(G,\chi)$ be the subset of $\Pi_{temp}(G,\chi)$ consisting of all the discrete series, and let $\Pi_{temp,ind}(G,\chi) = \Pi_{temp}(G,\chi)\smallsetminus\Pi_2(G,\chi)$. Let $\mathcal{C}(\Pi_2(G,\chi))$ (resp. $\mathcal{C}(\Pi_{temp,ind}(G,\chi))$) be the subspace of $\mathcal{C}(\Pi_{temp}(G,\chi))$ consisting of those $T : \pi \to T_\pi$ such that T is supported on the set $\Pi_2(G,\chi)$ (resp. $\Pi_{temp,ind}(G,\chi)$). Then any element $T \in \mathcal{C}(\Pi_{temp}(G,\chi))$ can be uniquely written as $T = T_1 + T_2$ with $T_1 \in \mathcal{C}(\Pi_2(G,\chi))$ and $T_2 \in \mathcal{C}(\Pi_{temp,ind}(G,\chi))$. In other words, we have

$$\mathcal{C}(\Pi_{temp}(G,\chi)) = \mathcal{C}(\Pi_2(G,\chi)) \oplus \mathcal{C}(\Pi_{temp,ind}(G,\chi)).$$

Under the matricial Paley-Wiener theorem, the subspaces $\mathcal{C}(\Pi_2(G,\chi))$ and $\mathcal{C}(\Pi_{temp,ind}(G,\chi))$ of $\mathcal{C}(\Pi_{temp}(G,\chi))$ allow us to define the corresponding subspaces of $\mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1})$. We define

$$\begin{aligned} {}^\circ\mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1}) &= \{f \in \mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1})\colon T_f \in \mathcal{C}(\Pi_2(G,\chi))\}, \\ \mathcal{C}_{ind}(Z_G(F)\backslash G(F),\chi^{-1}) &= \{f \in \mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1})\colon T_f \in \mathcal{C}(\Pi_{temp,ind}(G,\chi))\}. \end{aligned}$$

Then we have

$$\mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1}) = {}^\circ\mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1}) \oplus \mathcal{C}_{ind}(Z_G(F)\backslash G(F),\chi^{-1}).$$

It is easy to see that ${}^\circ\mathcal{C}(Z_G(F)\backslash G(F),\chi^{-1})$ is spanned by the matrix coefficients of all the discrete series of $G(F)$ with central character χ^{-1}.

2.4. **Quasi-characters.** If θ is a smooth function on $G_{reg}(F)$, invariant under $G(F)$-conjugation, we say it is a quasi-character if for every $x \in G_{ss}(F)$, there is a good neighborhood ω_x of 0 in $\mathfrak{g}_x(F)$, and for every $\mathcal{O} \in Nil(\mathfrak{g}_x)$, there exists $c_{\theta,\mathcal{O}}(x) \in \mathbb{C}$ such that

$$\theta(x\exp(X)) = \sum_{\mathcal{O}\in Nil(\mathfrak{g}_x)} c_{\theta,\mathcal{O}}(x)\hat{j}(\mathcal{O},X) \tag{2.2}$$

for every $X \in \omega_{x,reg}$. Here $\hat{j}(\mathcal{O},X)$ is the function on $\mathfrak{g}_{reg}(F)$ representing the Fourier transform of the nilpotent orbital integral (see Section 1.8 of [2]), $Nil(\mathfrak{g}_x)$ is the set of nilpotent orbits of $\mathfrak{g}_x(F)$, and we refer the reader to Section 3 of [15] for the definition of good neighborhood. The coefficients $c_{\theta,\mathcal{O}}(x)$ are called the germs of θ at x. We define

$$c_\pi(x) = \begin{cases} \frac{1}{|Nil_{reg}(\mathfrak{g}_x)|}\sum_{\mathcal{O}\in Nil_{reg}(\mathfrak{g}_x)} c_{\theta_\pi,\mathcal{O}}(x) & \text{if } Nil_{reg}(\mathfrak{g}_x) \neq \emptyset \\ 0 & \text{if } Nil_{reg}(\mathfrak{g}_x) = \emptyset. \end{cases}$$

Here $Nil_{reg}(\mathfrak{g}_x)$ is the set of regular nilpotent orbits of $\mathfrak{g}_x(F)$ (it is nonempty if and only if $G_x(F)$ is quasi-split). For any smooth finite length representation π of $G(F)$, the distribution character θ_π is a quasi-character.

Similarly, if θ is a smooth function on $\mathfrak{g}_{reg}(F)$, invariant under $G(F)$-conjugation. We say it is a quasi-character on $\mathfrak{g}(F)$ if for every $X \in \mathfrak{g}_{ss}(F)$, there exists an open G_X-invariant neighborhood $\omega_X \subset \mathfrak{g}_X(F)$ of 0, and for every $\mathcal{O} \in Nil(\mathfrak{g}_X(F))$, there exists $c_{\theta,\mathcal{O}}(X) \in \mathbb{C}$ such that

$$\theta(X+Y) = \sum_{\mathcal{O}\in Nil(\mathfrak{g}_X(F))} c_{\theta,\mathcal{O}}(X)\hat{j}(\mathcal{O},Y)$$

for every $Y \in \omega_{X,reg}$. As in the group case, we use $c_\theta(X)$ to denote the average of the germs associated to the regular nilpotent orbits.

2.5. **Strongly cuspidal functions.** We say a function $f \in \mathcal{C}(Z_G(F)\backslash G(F),\chi)$ is strongly cuspidal if for every proper parabolic subgroup $P = MU$ of G, and for every $x \in M(F)$, we have

$$\int_{U(F)} f(xu)\,\mathrm{d}u = 0. \tag{2.3}$$

We will denote by $\mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\chi)$ (resp. $C^\infty_{c,scusp}(Z_G(F)\backslash G(F),\chi)$) the subspace of strongly cuspidal functions in $\mathcal{C}(Z_G(F)\backslash G(F),\chi)$ (resp. $C_c^\infty(Z_G(F)\backslash G(F),\chi)$). It is easy to see that ${}^\circ\mathcal{C}(Z_G(F)\backslash G(F),\chi) \subset \mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\chi)$. Hence we have

$$\mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\chi) = {}^\circ\mathcal{C}(Z_G(F)\backslash G(F),\chi) \oplus \mathcal{C}_{ind,scusp}(Z_G(F)\backslash G(F),\chi)$$

where $\mathcal{C}_{ind,scusp}(Z_G(F)\backslash G(F),\chi)$ is the subspace of strongly cuspidal functions in $\mathcal{C}_{ind}(Z_G(F)\backslash G(F),\chi)$.

Similarly, we say a function $f \in C_c^\infty(\mathfrak{g}(F))$ is strongly cuspidal if for every proper parabolic subgroup $P = MU$, and for every $X \in \mathfrak{m}(F)$, we have

$$\int_{\mathfrak{u}(F)} f(X+Y)\,\mathrm{d}Y = 0.$$

We use $C^\infty_{c,scusp}(\mathfrak{g}(F))$ to denote the subspace of strongly cuspidal functions in $C_c^\infty(\mathfrak{g}(F))$.

We then define various objects associated to strongly cuspidal functions. Geometrically, for $f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\chi)$ (resp. $f \in C^\infty_{c,scusp}(\mathfrak{g}(F))$), one can define a quasi-character θ_f of $G(F)$ (resp. $\mathfrak{g}(F)$) via the weighted orbital integral (see Section 5.2 of [2] for details). Spectrally, let $\mathcal{X}(G,\chi^{-1})$ (resp. $\mathcal{X}_{ell}(G,\chi^{-1})$) be the set of virtual tempered representations (resp. elliptic representations) of $G(F)$ with central character χ^{-1} defined in Section 2.7 of [2]. As in Section 5.4 of loc. cit., for $\pi \in \mathcal{X}(G,\chi^{-1})$, we can define a map

$$f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\chi) \mapsto \theta_f(\pi) \in \mathbb{C}$$

via the weighted character (this map was denoted by $f \mapsto \hat{\theta}_f(\pi)$ in loc. cit.). We refer the reader to Section 5 of [2] for basic properties of strongly cuspidal functions.

2.6. **Some local representation theory of U_{2n} and GU_{2n}.** In this subsection, we recall some results of the local representation theory of the unitary groups and the unitary similitude groups. For $i=1,2$, fix $\varepsilon_i \in F^\times$ with $\eta_{E/F}(\varepsilon_i) = (-1)^{i-1}$ as before. We start with the unitary group case. Let $\Pi_{irr,temp}(\mathrm{U}_{2n}) = \Pi_{irr,temp}(\mathrm{U}(J_{2n,\varepsilon_1})) \cup \Pi_{irr,temp}(\mathrm{U}(J_{2n,\varepsilon_2}))$ be the set of isomorphism classes of irreducible tempered representations of $\mathrm{U}(J_{2n,\varepsilon_1})(F)$ and $\mathrm{U}(J_{2n,\varepsilon_2})(F)$. The following theorem follows from the endoscopic classification of unitary groups in [10] and [9].

Theorem 2.1 ([10], [9])**.** *$\Pi_{irr,temp}(\mathrm{U}_{2n})$ is a disjoint union of finite sets (i.e. the local tempered Vogan L-packets)*

$$\Pi_{irr,temp}(\mathrm{U}_{2n}) = \cup_\phi \Pi_\phi$$

where ϕ runs over all the tempered L-parameters of $\mathrm{U}_{2n}(F)$ and $\Pi_\phi = \Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_1})) \cup \Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_2}))$ consists of a finite number of tempered representations with $\Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_i})) \subset \Pi_{irr,temp}(\mathrm{U}(J_{2n,\varepsilon_i}))$ such that the following conditions hold.

(1) For all ϕ, the distribution character

$$\theta_{\Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_i}))} := \sum_{\pi_{\varepsilon_i} \in \Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_i}))} \theta_{\pi_{\varepsilon_i}}$$

is stable for $i=1,2$, i.e. $\theta_{\Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_i}))}(g) = \theta_{\Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_i}))}(g')$ for all $g, g' \in \mathrm{U}(J_{2n,\varepsilon_i})_{reg}(F)$ belonging to the same stable conjugacy class.

(2) Let ψ be any generic character of a maximal unipotent subgroup of $\mathrm{U}(J_{2n,\varepsilon_1})$ (up to conjugation, there are two such characters). For all ϕ, the L-packet $\Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_1}))$ contains a unique generic representation with respect to ψ (note that $\mathrm{U}(J_{2n,\varepsilon_2})(F)$ is not quasi-split).

(3) For a pair $g_i \in \mathrm{U}(J_{2n,\varepsilon_i})(F)$, we write $g_1 \leftrightarrow g_2$ if they have the same characteristic polynomial (i.e. they are in the same stable conjugacy class). Then for any pair $g_i \in \mathrm{U}(J_{2n,\varepsilon_i})_{reg}(F)$ with $g_1 \leftrightarrow g_2$, we have

$$\theta_{\Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_1}))}(g_1) = -\theta_{\Pi_\phi(\mathrm{U}(J_{2n,\varepsilon_2}))}(g_2).$$

Remark that Item 2 of the theorem was proved in [8] by assuming the endoscopic identity. The endoscopic identity for unitary groups has been proved in [10] and [9].

Now we consider the unitary similitude group case. Recall that for $\varepsilon \in F^\times$, $\mathrm{GU}_{2n,\varepsilon} = \mathrm{GU}(J_{2n,\varepsilon})$ and $\mathrm{U}_{2n,\varepsilon} = \mathrm{U}(J_{2n,\varepsilon})$. We have

$$\mathrm{GU}_{2n,\varepsilon}(F)/Z_{\mathrm{GU}_{2n,\varepsilon}}(F)\mathrm{U}_{2n,\varepsilon}(F) \cong \lambda(\mathrm{GU}_{2n,\varepsilon}(F))/\lambda(Z_{\mathrm{GU}_{2n,\varepsilon}}(F)) = F^\times/\mathrm{Im}(N_{E/F}) \cong \mathbb{Z}/2\mathbb{Z},$$

$$\mathrm{Hom}(\mathrm{GU}_{2n,\varepsilon}(F)/Z_{\mathrm{GU}_{2n,\varepsilon}}(F)\mathrm{U}_{2n,\varepsilon}(F), \mathbb{C}^\times) = \{1, \lambda_{E/F}\}$$

where $\lambda_{E/F} = \eta_{E/F} \circ \lambda$ is a character of $\mathrm{GU}_{2n,\varepsilon}(F)$.

Lemma 2.2 (Corollary 6.7 [24])**.** *Let π_ε be an irreducible smooth representation of $\mathrm{GU}_{2n,\varepsilon}(F)$ and let $\pi_\varepsilon|_{\mathrm{U}_{2n,\varepsilon}(F)}$ be the restriction of π_ε to $\mathrm{U}_{2n,\varepsilon}(F)$. Then $\pi_\varepsilon|_{\mathrm{U}_{2n,\varepsilon}(F)}$ is multiplicity-free. Moreover, $\pi_\varepsilon|_{\mathrm{U}_{2n,\varepsilon}(F)}$ is irreducible if and only if $\pi_\varepsilon \not\cong \pi_\varepsilon \otimes \lambda_{E/F}$. If $\pi_\varepsilon \cong \pi_\varepsilon \otimes \lambda_{E/F}$, then $\pi_\varepsilon|_{\mathrm{U}_{2n,\varepsilon}(F)} = \pi \oplus \pi \circ \mathrm{Ad}(g)$*

for some irreducible representation π *of* $\mathrm{U}_{2n,\varepsilon}(F)$*, where* $g \in \mathrm{GU}_{2n,\varepsilon}(F)$ *with* $\lambda(g) \notin \mathrm{Im}(N_{E/F})$*.*

Lemma 2.3 (Corollary 6.4 [24])**.** *If* π *is an irreducible smooth representation of* $\mathrm{U}_{2n,\varepsilon}(F)$*, then there exists an irreducible smooth representation* π_ε *of* $\mathrm{GU}_{2n,\varepsilon}(F)$*, which is unique up to twisting by the characters* $\chi \circ \lambda$ *where* χ *is any character of* $F^\times$*, such that* π *is a direct summand of* $\pi_\varepsilon|_{\mathrm{U}_{2n,\varepsilon}(F)}$*.*

Lemma 2.4 (Lemma 6.9 [24])**.** *Suppose that* π_ε *is an irreducible smooth unitary representation of* $\mathrm{GU}_{2n,\varepsilon}(F)$*. Then* π_ε *is a discrete series if and only if its restriction to* $\mathrm{U}_{2n,\varepsilon}(F)$ *is a discrete series (not necessarily irreducible). The same is true for tempered representations.*

Let $\Pi_{irr,temp}(\mathrm{GU}_{2n}) = \Pi_{irr,temp}(\mathrm{GU}(J_{2n,\varepsilon_1})) \cup \Pi_{irr,temp}(\mathrm{GU}(J_{2n,\varepsilon_2}))$ be the set of isomorphism classes of irreducible tempered representations of $\mathrm{GU}(J_{2n,\varepsilon_1})(F)$ and $\mathrm{GU}(J_{2n,\varepsilon_2})(F)$. In order to prove our main theorems, we need to assume that the following conjecture holds. This conjecture is the endoscopic classification of the unitary similitude groups. By the endoscopic classification of the unitary groups ([10], [9]), this conjecture is expected from Xu's work [24] on the reduction from the similitude classical groups to classical groups.

Conjecture 2.5. $\Pi_{irr,temp}(\mathrm{GU}_{2n})$ *is a disjoint union of finite sets (i.e. the local tempered Vogan L-packets)*

$$\Pi_{irr,temp}(\mathrm{GU}_{2n}) = \cup_\phi \Pi_\phi$$

where ϕ *runs over all the tempered L-parameters of* $\mathrm{GU}_{2n}(F)$ *and* $\Pi_\phi = \Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_1})) \cup \Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_2}))$ *consists of a finite number of tempered representations with* $\Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_i})) \subset \Pi_{irr,temp}(\mathrm{GU}(J_{2n,\varepsilon_i}))$ *such that the following conditions hold.*

(1) For all ϕ*, the distribution character*

$$\theta_{\Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_i}))} := \sum_{\pi_{\varepsilon_i} \in \Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_i}))} \theta_{\pi_{\varepsilon_i}}$$

is stable for $i = 1, 2$*, i.e.* $\theta_{\Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_i}))}(g) = \theta_{\Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_i}))}(g')$ *for all* $g, g' \in \mathrm{GU}(J_{2n,\varepsilon_i})_{reg}(F)$ *belonging to the same stable conjugacy class.*

(2) For all ϕ*, the L-packet* $\Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_1}))$ *contains a unique generic representation (note that* $\mathrm{GU}(J_{2n,\varepsilon_2})(F)$ *is not quasi-split).*

(3) For $g_i \in \mathrm{GU}(J_{2n,\varepsilon_i})(F)$*, we write* $g_1 \leftrightarrow g_2$ *if they have the same characteristic polynomials (i.e. they are in the same stable conjugacy class). Then*

$$\theta_{\Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_1}))}(g_1) = -\theta_{\Pi_\phi(\mathrm{GU}(J_{2n,\varepsilon_2}))}(g_2)$$

for all pairs $g_i \in \mathrm{GU}(J_{2n,\varepsilon_i})_{reg}(F)$ *with* $g_1 \leftrightarrow g_2$*.*

3. The model $(G_\varepsilon, H_\varepsilon)$

3.1. **The spherical pair** $(G_\varepsilon, H_\varepsilon)$**.** Let $(G,H)=(G,H_0U)$ be the pair $(G_\varepsilon, H_{0,\varepsilon}U_\varepsilon)$ defined in Section 1 for some $\varepsilon\in F^\times$, and let $G_0=M_\varepsilon$. Then the spherical pair (G,H) can be viewed as the Whittaker induction of the spherical pair (G_0,H_0). We will use $\omega\otimes\xi$ to denote the character $\omega_\varepsilon\otimes\xi_\varepsilon$. For simplicity, we omit the subscript ε here. We say a parabolic subgroup $\bar{Q}$ of G is good if $H\bar{Q}$ is a Zariski open subset of G. This is equivalent to say that $H(F)\bar{Q}(F)$ is open in $G(F)$ under the analytic topology. The proof of the next proposition is very similar to the GR model case (Proposition 4.2 of [18]), so we will skip it here. The only thing we want to point out is that the proposition will only hold for the unitary similitude group case as we will have two open Borel orbits (which correspond to $F^\times/\mathrm{Im}(N_{E/F})$) for the unitary group case.

Proposition 3.1. *(1) There exist minimal parabolic subgroups of G that are good and they are all conjugated to each other by some elements in $H(F)$. If $\bar{P}_{min}=M_{min}\bar{U}_{min}$ is a good minimal parabolic subgroup, we have $H\cap\bar{U}_{min}=\{1\}$ and the complement of $H(F)\bar{P}_{min}(F)$ in $G(F)$ has zero measure.*

(2) A parabolic subgroup $\bar{Q}$ of G is good if and only if it contains a good minimal parabolic subgroup.

(3) Let $\bar{P}_{min}=M_{min}\bar{U}_{min}$ be a good minimal parabolic subgroup and let $A_{min}(F)=A_{M_{min}}(F)$ be the maximal split torus of the center of $M_{min}(F)$. Set

$$A_{min}^+=\{a\in A_{min}(F)\colon |\alpha(a)|\geq 1 \text{ for all } \alpha\in\Psi(A_{min},\bar{P}_{min})\}.$$

Then we have

(a) $\sigma_0(h)+\sigma_0(a)\ll\sigma_0(ha)$ for all $a\in A_{min}^+$, $h\in H(F)$.

(b) $\sigma(h)\ll\sigma(a^{-1}ha)$ and $\sigma_0(h)\ll\sigma_0(a^{-1}ha)$ for all $a\in A_{min}^+$, $h\in H(F)$.

(4) (1),(2) and (3) also hold for the pair (G_0,H_0).

By the proposition above, $X=H\backslash G$ is a spherical variety of G and $X_0=H_0\backslash G_0$ is a spherical variety of G_0. Let $\bar{P}_0=M_0\bar{U}_0$ be a good minimal parabolic subgroup of G_0 and let $A_0(F)=A_{M_0}(F)$. Set

$$A_0^+=\{a\in A_0(F)\colon |\alpha(a)|\geq 1,\ \forall\alpha\in\Psi(A_0,\bar{P}_0)\}.$$

By a similar argument as in the Ginzburg–Rallis model case (Proposition 4.4 of [18]), we can prove the weak Cartan decomposition of X and X_0.

Proposition 3.2. *(1) There exists a compact subset $\mathcal{K}_0\subset G_0(F)$ such that $G_0(F)=H_0(F)A_0^+\mathcal{K}_0$.*

(2) There exists a compact subset $\mathcal{K}\subset G(F)$ such that $G(F)=H(F)A_0^+\mathcal{K}$.

3.2. **Some estimates.** In this subsection, we are going to state several estimates for various integrals which will be used in later sections. The proofs of these estimates are very similar to the GR model case (Section 4.3 and 4.4 of [20]).

Lemma 3.3. *(1) There exist $\epsilon > 0$ and $d > 0$ such that the integrals*

$$\int_{Z_{H_0}(F)\backslash H_0(F)} \Xi^{G_0}(h_0)e^{\epsilon\sigma_0(h_0)}\,\mathrm{d}h_0, \quad \int_{Z_H(F)\backslash H(F)} \Xi^{G}(h)\sigma_0(h)^{-d}\,\mathrm{d}h$$

are absolutely convergent.

(2) For all $\delta > 0$, there exists $\epsilon > 0$ such that the integral

$$\int_{Z_H(F)\backslash H(F)} \Xi^{G}(h)e^{\epsilon\sigma_0(h)}(1+|\iota(h)|)^{-\delta}\,\mathrm{d}h$$

is absolutely convergent. Here $\iota\colon H(F) \to F$ is a homomorphism defined by

$$\iota(\begin{pmatrix} I_2 & X & Y \\ & I_2 & X^* \\ & & I_2 \end{pmatrix}\begin{pmatrix} g & & \\ & h & \\ & & \lambda(h)g^* \end{pmatrix}) = \mathrm{tr}_{E/F}(\mathrm{tr}(X)).$$

In particular, we have $\xi(h) = \psi(\iota(h))$ for all $h \in H(F)$.

Proof. The proof is very similar to the GR model case in Lemma 4.3.1 of [20], we will skip it here. □

Let $C \subset G(F)$ be a compact subset with non-empty interior. Define the function $\Xi_C^{H\backslash G}(x) = vol_{H\backslash G}(xC)^{-1/2}$ for $x \in H(F)\backslash G(F)$. If C' is another compact subset with non-empty interior, then $\Xi_C^{H\backslash G}(x) \sim \Xi_{C'}^{H\backslash G}(x)$ for all $x \in H(F)\backslash G(F)$. We will only use the function $\Xi_C^{H\backslash G}$ for majorization. From now on, we fix a particular C, and set $\Xi^{H\backslash G} = \Xi_C^{H\backslash G}$. The proof of the next two propositions follows from the exact same arguments as in GR model case (Proposition 4.4.1 and Lemma 4.4.2 of [20]), we will skip it here.

Proposition 3.4. *(1) There exists $d > 0$ such that the integral*

$$\int_{H(F)\backslash G(F)} \Xi^{H\backslash G}(x)^2\sigma_{H\backslash G}(x)^{-d}\,\mathrm{d}x$$

is absolutely convergent. Here $\sigma_{H\backslash G}(x) := \inf_{h\in H(F)}\sigma(hx)$ for $x \in H(F)\backslash G(F)$.

(2) For all $d > 0$, there exists $d' > 0$ such that

$$\int_{Z_H(F)\backslash H(F)} \Xi^{G}(hx)\sigma_0(hx)^{-d'}\,\mathrm{d}h \ll \Xi^{H\backslash G}(x)\sigma_{H\backslash G}(x)^{-d}$$

for all $x \in H(F)\backslash G(F)$.

Proposition 3.5. *Let $\bar{Q} = M_Q\bar{U}_Q$ be a good parabolic subgroup of G. Let $H_{\bar{Q}} = H \cap \bar{Q}$, and let $G_{\bar{Q}} = \bar{Q}/\bar{U}_Q$ be the reductive quotient of $\bar{Q}$. Then*

(1) $H_{\bar{Q}} \cap \bar{U}_Q = \{1\}$, hence we can view $H_{\bar{Q}}$ as a subgroup of $G_{\bar{Q}}$. We also have $\delta_{\bar{Q}}(h_{\bar{Q}}) = \delta_{H_{\bar{Q}}}(h_{\bar{Q}})$ for all $h_{\bar{Q}} \in H_{\bar{Q}}(F)$.

(2) There exists $d > 0$ *such that the integral*

$$\int_{Z_H(F)\backslash H_{\bar{Q}}(F)} \Xi^{G_{\bar{Q}}}(h_{\bar{Q}})\sigma_0(h_{\bar{Q}})^{-d}\delta_{H_{\bar{Q}}}(h_{\bar{Q}})^{1/2}\,\mathrm{d}h_{\bar{Q}}$$

is absolutely convergent where $\mathrm{d}h_{\bar{Q}}$ *is a left Haar measure on* $Z_H(F)\backslash H_{\bar{Q}}(F)$.

4. The trace formula and the multiplicity formula

4.1. **The distribution $I_{geom}(f)$.** As in the previous section, we will use (G, H) to denote the pair $(G_\varepsilon, H_\varepsilon)$ and $\omega \otimes \xi$ to denote the character $\omega_\varepsilon \otimes \xi_\varepsilon$. Given $f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F), \eta^{-1})$, we have associated the quasi-character θ_f on $G(F)$ in Section 2. Let $\mathcal{T}_{ell}(H_0)$ be a set of representatives of conjugacy classes of maximal elliptic tori in H_0. In both cases (i.e. the quasi-split case and the non quasi-split case), there is a natural bijection between $\mathcal{T}_{ell}(H_0)$ and the set of all the quadratic extensions of F. For all $T \in \mathcal{T}_{ell}(H_0)$ and $t \in T(F)_{reg}$, if the torus corresponds to E, then G_t is isomorphic to $G(U_3 \times U_3)$ where GU_3 is the unitary similitude group in three variables and $G(U_3 \times U_3) = \{(g_1, g_2) \in \mathrm{GU}_3 \times \mathrm{GU}_3 |\ \lambda(g_1) = \lambda(g_2)\}$ (λ is the similitude character). If the torus T corresponds to a quadratic extension F' of F with $F' \neq E$, let $E' = F' \otimes_F E$ which is a quadratic extension of F'. Then the centralizer G_t is isomorphic to the subgroup of the similitude group of three variables with respect to the quadratic extension E'/F' (denoted by $\mathrm{GU}_3(E'/F')$) consisting of elements whose similitude belongs to $F^\times$, i.e. $G_t(F) = \{g \in \mathrm{GU}_3(E'/F')|\ \lambda(g) \in F^\times\}$. In both cases, the Lie algebra of $G_t(F)$ has a unique regular nilpotent orbit which will be denoted by $\mathcal{O}_t$. We then define

$$c_f(t) := c_{\theta_f}(t) = c_{\theta_f,\mathcal{O}_t}(t).$$

Definition 4.1. *Define the geometric side of the trace formula to be*

$$I_{geom}(f) := c_f(1) + \sum_{T\in\mathcal{T}_{ell}(H_0)} |W(H_0,T)|^{-1}\mathrm{vol}(T(F)/Z_G(F))^{-1}\int_{T(F)/Z_G(F)} D^H(t)c_f(t)\omega(t)\,\mathrm{d}t$$

where $c_f(1) = c_{\theta_f}(1)$ *is the regular germ of* θ_f *at* 1 *(which is always equal to* 0 *if the group is not quasi-split) and* $W(H_0, T)$ *is the Weyl group. By a similar argument as in the GR case (Proposition 5.2 of* [17]*), we know that the integral above is absolutely convergent.*

4.2. **The distribution $I(f)$ and $I_{spec}(f)$.** For $f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F), \eta^{-1})$, define the function $I(f, \cdot)$ on $H(F)\backslash G(F)$ to be

$$I(f,g) = \int_{Z_H(F)\backslash H(F)} f(g^{-1}hg)\omega\otimes\xi(h)\,\mathrm{d}h.$$

By Lemma 3.3, the above integral is absolutely convergent. Then by the same argument as in the GGP case (Proposition 8.1.1 of [2]) and the GR case (Appendix B of [20]), we can show that the integral

$$I(f) := \int_{H(F)\backslash G(F)} I(f,g)\,\mathrm{d}g \tag{4.1}$$

is absolutely convergent for all $f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\eta^{-1})$, and it defines a continuous linear form

$$\mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\eta^{-1}) \to \mathbb{C}\colon f \to I(f).$$

$I(f)$ will be the distribution in our trace formula.

Remark 4.2. *As in the GGP case and the GR case, although the integral* (4.1) *defining $I(f)$ is absolutely convergent, the double integral*

$$\int_{H(F)\backslash G(F)} \int_{Z_H(F)\backslash H(F)} f(g^{-1}hg)\omega\otimes\xi(h)\,\mathrm{d}h\,\mathrm{d}g$$

is not absolutely convergent. As a result, in the proof of the geometric side of the trace formula, we need to introduce truncation functions on $H(F)\backslash G(F)$.

We then define the spectral side of the trace formula to be

$$I_{spec}(f) = \int_{\mathcal{X}(G,\eta)} D(\pi)\theta_f(\pi)m(\bar{\pi})\,\mathrm{d}\pi. \tag{4.2}$$

Here $\bar{\pi}$ is the complex conjugation of π which is isomorphic to the dual representation of π since π is unitary. We refer the reader to Section 2.7 of [2] for the definitions of $D(\pi)$ and the measure $\mathrm{d}\pi$. Now we are ready to state the trace formula.

Theorem 4.3. *For all $f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\eta^{-1})$, we have*

$$I_{spec}(f) = I(f) = I_{geom}(f).$$

The spectral expansion will be proved in Section 6, while the geometric expansion will be proved in the next subsection.

To end this subsection, we define the Lie algebra analog of the distribution $I(f)$ in the trace formula. This will be used in the proof of the geometric expansion. Denote $\mathfrak{g}'(F)$ (resp. $\mathfrak{h}'(F)$, $\mathfrak{h}_0'(F)$) to be the subspace of $\mathfrak{g}(F)$ (resp. $\mathfrak{h}(F)$, $\mathfrak{h}_0(F)$) consisting of elements of trace zero. Then $\mathfrak{g}(F) = \mathfrak{g}'(F)\oplus\mathfrak{z}_{\mathfrak{g}}(F)$, $\mathfrak{h}(F) = \mathfrak{h}'(F)\oplus\mathfrak{z}_{\mathfrak{g}}(F)$ and $\mathfrak{h}_0(F) = \mathfrak{h}_0'(F)\oplus\mathfrak{z}_{\mathfrak{g}}(F)$. For $\phi \in C_{c,scusp}^{\infty}(\mathfrak{g}'(F))$, we define

$$I(\phi,g) = \int_{\mathfrak{h}_0'(F)}\int_{\mathfrak{u}(F)} \phi(g^{-1}(X+Y)g)\xi(\exp(Y))\,\mathrm{d}Y\,\mathrm{d}X, \quad I(\phi) = \int_{H(F)\backslash G(F)} I(\phi,g)\,\mathrm{d}g.$$

As in the group case, the integral defining $I(\phi)$ is absolutely convergent.

4.3. **The proof of the geometric expansion.** In this subsection, we prove the geometric side of the trace formula. The idea of the proof is the same as the GGP case in [15] and [2], while all the computations are very similar to the GR case in [17]. As a result, we will only give a sketch of the proof without providing details.

First, by the standard argument as in the GGP case, once we have proved the spectral side of the trace formula (this will be done in Section 6), we only need to prove the geometric side for compactly supported functions (i.e. for $f \in C_{c,scusp}^{\infty}(Z_G(F)\backslash G(F),\eta^{-1})$). To be specific, once we have proved the spectral expansion, we can prove the following proposition which is an analog of Proposition 11.5.1 of [2] for the current model.

Proposition 4.4. *The distribution $I(f)$ only depends on the quasi-character θ_f. Moreover, there exists a continuous linear form $J(\cdot)$ on $QC(G(F),\eta^{-1})$, the space of quasi-characters of $G(F)$ with central character η^{-1}, such that*

- $J(\theta_f) = I(f) - I_{geom}(f)$ *for all* $f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F),\eta^{-1})$.
- $J(\cdot)$ *is supported on* $G(F)_{ell}$.

We refer the reader to Section 4.1 of [2] *for the definition of the topology on* $QC(G(F),\eta^{-1})$.

Proof. The proof is the same as Proposition 11.5.1 of [2], we will skip it here. □

Combining the above proposition with Corollary 5.7.2 of [2], we know that in order to show the linear form $J(\cdot)$ is identically zero, we only need to show that $J(\theta_f) = 0$ for $f \in C_{c,scusp}^{\infty}(Z_G(F)\backslash G(F),\eta^{-1})$. Hence it is enough to prove the geometric side for compactly supported functions. Note that although Corollary 5.7.2 of [2] only considered the case when A_G is trivial, it can be easily extended to the general case once we have fixed a central character on A_G.

Next we reduce the proof of the geometric expansion to the case when the characters χ_E, χ_F and η are all trivial. Up to twisting the test functions by the character $\chi_F(\lambda(\cdot))$ we may assume that the character χ_F is trivial. In this case, we have $\eta = \chi_E^2$ and in particular we have $\eta(-1) = 1$. By replacing f by $f \cdot 1_{\Omega\cdot Z_G(F)}$ where Ω is an G-invariant open subset of $G(F)$ (here G-invariant means that Ω is invariant under conjugation), we may assume that f is supported on $Z_G(F)\cdot\Omega$ where Ω is a small G-invariant neighborhood of a semisimple element $x \in G_{ss}(F)$ of G.

If x is not conjugate to an element of $H_0(F)$, then by taking Ω small we will have $I(f) = I_{geom}(f) = 0$ and this proves the geometric expansion. So we may assume that $x \in H_{0,ss}(F)$. Let Ω_E be a small open compact subset of $E^{\times}$ containing 1 such that the characters η and χ_E are trivial on Ω_E. Since the multiset of the eigenvalues of x is of the form (a,a,a,b,b,b) for some $a,b \in \bar{F}$, we may choose Ω small enough such that the following property holds.

(1) If $x_1 = zx_2$ for some $x_1, x_2 \in \Omega$ and $z \in Z_G(F)$, then $z = aI_6$ with $a \in \Omega_E$ or $-a \in \Omega_E$.

Now we define a function f' on $G(F)$ to be

$$f'(zx_0) = f(x_0),\ x_0 \in \Omega, z \in Z_G(F).$$

By the Condition (1) above and the fact that $\eta(-1) = 1$, we know that f' is well defined and it is a strongly cuspidal function on $G(F)$ with trivial central character. As a result, we can define

$$I(f',g) = \int_{Z_H(F)\backslash H(F)} f'(g^{-1}hg)\xi(h)\,\mathrm{d}h,\ I(f') = \int_{H(F)\backslash G(F)} I(f',g)\,\mathrm{d}g.$$

Recall that we have defined

$$I(f,g)=\int_{Z_H(F)\backslash H(F)} f(g^{-1}hg)\omega\otimes\xi(h)dh.$$

Since Ω is G-invariant, for any $g\in G(F)$, we can choose a fundamental domain $\Omega(H,g)$ of $Z_H(F)\backslash H(F)$ such that $(g^{-1}\Omega(H,g)g)\cap(\Omega Z_G(F))\subset\Omega$. Moreover, by choosing Ω small we may assume that $\omega\otimes\xi(h)=\xi(h)\omega(x)$ for all $h\in\Omega\cap H(F)$. By the definition of f' together with the fact that both f and f' are supported on $\Omega Z_G(F)$, we have

$$f(g^{-1}hg)\omega\otimes\xi(h)=\omega(x)\xi(h)f'(g^{-1}hg)$$

for all $h\in\Omega(H,g)$. This implies that $I(f,g)=\omega(x)I(f',g)$ and hence $I(f)=\omega(x)I(f')$. Similarly we can also show that $I_{geom}(f)=\omega(x)I(f')$. As a result, we have reduced the proof of the geometric expansion to the case when the characters χ_E,χ_F and η are all trivial.

Remark 4.5. *The above argument is very similar to the argument for the GR case in Proposition 5.6 of* [17]. *However, as pointed out by the referee, the argument in the loc. cit. is incorrect because the set g_0X in loc. cit. is not invariant under $G(F)$-conjugation. In the above argument, we fixed this error by choosing the neighborhood ω which is invariant under conjugation. Our argument here also works for the GR case and hence it fixes the error in the proof of Proposition 5.6 of* [17].

The next step is to study the distribution $I(\phi)$ for the Lie algebra case (i.e. for $\phi\in C_{c,scusp}^{\infty}(\mathfrak{g}'(F))$). The goal is to express $I(\phi)$ in terms of $\theta_{\hat{\phi}}=\hat{\theta}_{\phi}$ where $\hat{\phi}$ is the Fourier transform of ϕ. In order to do this, we first need to introduce a sequence of truncation functions $\kappa_N\in C_c^{\infty}(H(F)\backslash G(F))$ $(N\geq 1)$ whose definition is similar to the GR case. To be specific, for each $g\in G(F)$, let $g=umk$ be the Iwasawa decomposition of g with $m=m(g_1,h)\in M(F)$ (recall that $m(g_1,h)=diag(g_1,h,\lambda(h)g_1^{\ast})$ for $g_1\in \mathrm{GL}_2(E)$ and $g_2\in \mathrm{GU}_2(F)$). We define (here $\sigma_{\mathrm{GL}_2(E)}$ is a log-norm on $\mathrm{GL}_2(E)$)

$$\kappa_N(g)=\begin{cases}1, & \text{if } \sigma_{\mathrm{GL}_2(E)}(g_1h^{-1})\leq N;\\ 0, & \text{otherwise.}\end{cases}$$

For $N\geq 1$, we define

$$I_N(\phi)=\int_{H(F)\backslash G(F)}\kappa_N(g)I(\phi,g)\,\mathrm{d}g,\ \phi\in C_{c,scusp}^{\infty}(\mathfrak{g}'(F))$$

where

$$I(\phi,g)=\int_{\mathfrak{h}_0'(F)}\int_{\mathfrak{u}(F)}\phi(g^{-1}(X+Y)g)\xi(\exp(Y))\,\mathrm{d}Y\,\mathrm{d}X.$$

Recall that $\mathfrak{g}'(F)$ (resp. $\mathfrak{h}'(F)$, $\mathfrak{h}_0'(F)$) is the subspace of $\mathfrak{g}(F)$ (resp. $\mathfrak{h}(F)$, $\mathfrak{h}_0(F)$) consisting of elements of trace zero. Then $I(\phi)=\lim_{N\to\infty}I_N(\phi)$ and it is enough to consider $I_N(\phi)$.

Next we study the slice representation which is the conjugation action of $H(F)/Z_G(F)$ on the space $\Xi+\mathfrak{h}^{\perp}(F)$. Here $\mathfrak{h}^{\perp}$ is the orthogonal complement of $\mathfrak{h}$ in $\mathfrak{g}$ with respect to the Killing form and $\Xi = \begin{pmatrix} 0 & 0 & 0 \\ I_2 & 0 & 0 \\ 0 & -w_2 A_{\varepsilon} & 0 \end{pmatrix}$ is an element in $\bar{\mathfrak{u}}(F)$ associated to the character ξ of $U(F)$.

Lemma 4.6. *There exists a Zariski open subset of $\Xi + \mathfrak{h}^{\perp}(F)$, denoted by $\Xi + \mathfrak{h}^{\perp,0}(F)$, such that the following hold.*

1. *The group $Z_G(F)\backslash H(F)$ acts freely on the set $\Xi + \mathfrak{h}^{\perp,0}(F)$.*
2. *For $X, Y \in \Xi + \mathfrak{h}^{\perp,0}(F)$, X and Y are conjugated to each other in $G(F)$ if and only if they are conjugated to each other by an element of $H(F)$.*

Proof. This is a direct consequence of the computation in the Ginzburg-Rallis model case and the fact that $H^1(F, Z_G) = 1$. To be specific, by Lemma 8.2 and Lemma 8.4 of [17], there exists a Zariski open subset of $\Xi + \mathfrak{h}^{\perp}(F)$, denoted by $\Xi + \mathfrak{h}^{\perp,0}(F)$, such that

1. the group $Z_G(F)\backslash H(F)$ acts freely on the set $\Xi + \mathfrak{h}^{\perp,0}(F)$;
2. for $X, Y \in \Xi + \mathfrak{h}^{\perp,0}(F)$, X and Y are conjugated to each other in $G(F)$ if and only if they are conjugated to each other by an element of $H(E)$.

So we only need to show that for $X, Y \in \Xi+\mathfrak{h}^{\perp,0}(F)$, if there exists $h \in H(E)$ such that $h^{-1}Xh = Y$, then there exists $h_0 \in H(F)$ such that $(h_0)^{-1}Xh_0 = Y$. Let σ be the involution on $H(E)$ induced by the nontrivial element in $Gal(E/F)$. Since both X and Y are fixed by σ, together with condition (1) above, we know that $z = h\sigma(h)^{-1} \in Z_G(E)$. Since $H^1(F, Z_G) = 1$, we can find an element $z_0 \in Z_G(E)$ such that $z = z_0\sigma(z_0)^{-1}$. Then we just need to let $h_0 = hz_0^{-1}$. □

Remark 4.7.
1. *This lemma will fail for the unitary group case (i.e. the model $(G_{1,\varepsilon}, H_{1,\varepsilon})$) because the center of unitary group has nontrivial Galois cohomology. It is one of the reasons why we cannot prove the trace formula for the unitary group case (see Section 1.2).*
2. *This lemma is the key ingredient in the proof of the geometric expansion. The rest part of the proof follows from a standard argument (although it is very technical). In general, for any model (G, H) satisfies this lemma, if one can prove the spectral side of the trace formula, the geometric side of the trace formula can be proved by the same method as in this section (of course one needs to overcome some technical issues pertaining to the particular features of the model but they are not essential obstacles).*

The above lemma implies that we have an injective analytic morphism

$$(\Xi + \mathfrak{h}^{\perp,0})/H(F) \longrightarrow \coprod_{T\in\mathcal{T}(G)} \mathfrak{t}'(F)/W(G,T). \tag{4.3}$$

where $\mathcal{T}(G)$ is a set of representatives of conjugacy classes of maximal tori of $G(F)$, $W(G,T)$ is the Weyl group and $\mathfrak{t}'(F) = \mathfrak{t}(F) \cap \mathfrak{g}'(F)$. For each $T \in \mathcal{T}(G)$, let $\mathfrak{t}^0(F)/W(G,T)$ be the image of the map above. By the same argument as in the Ginzburg-Rallis model case (Remark 8.6 and Lemma 8.7 of [17]), we know that $\mathfrak{t}^0(F)$ is an open subset of $\mathfrak{t}'(F)$ and the Jacobian of this map is the square root of the Weyl determinant. For $\mathcal{T} \in T(G)$, we can fix a locally analytic map

$$\mathfrak{t}^0(F) \to \Xi + \mathfrak{h}^{\perp,0} : Y \to Y_\Sigma \tag{4.4}$$

such that the following diagram commutes:

$$\begin{array}{ccccc} \Xi + \mathfrak{h}^{\perp,0} & & \longrightarrow & & \mathfrak{t}^0(F)/W(G,T) \\ & \nwarrow & & \nearrow & \\ & & \mathfrak{t}^0(F) & & \end{array}$$

Then we can also find a map $Y \to \gamma_Y$ such that $Y_\Sigma = \gamma_Y^{-1} Y \gamma_Y$.

After studying the slice representation, by the same argument as in Section 8.6 of [17], we can rewrite $I_N(\phi)$ as an integral of some weighted orbital integrals of $\hat{\phi}$ where $\hat{\phi}$ is the Fourier transform of ϕ. To be specific, we have

$$\begin{aligned} I_N(\phi) &= \int_{H(F)\backslash G(F)} \kappa_N(g) \int_{\mathfrak{h}_0'(F)} \int_{\mathfrak{u}(F)} \phi(g^{-1}(X+Y)g)\xi(\exp(Y))\,\mathrm{d}Y\,\mathrm{d}X\,\mathrm{d}g \\ &= \int_{H(F)\backslash G(F)} \kappa_N(g) \int_{\Xi+\mathfrak{h}^\perp(F)} \hat{\phi}(g^{-1}Xg)\,\mathrm{d}X\,\mathrm{d}g. \end{aligned}$$

Here the equation on the second line follows from the Fourier transform on the Lie algebra (as in Lemma 8.1 of [17]). By the discussion of the slice representation above, we can rewrite the inner integral as

$$\sum_{T\in\mathcal{T}(G)} \mid W(G,T) \mid^{-1} \int_{Z_H(F)\backslash H(F)} \int_{\mathfrak{t}^0(F)} \hat{\phi}(g^{-1}y^{-1}\gamma_Y^{-1}Y\gamma_Y yg) D^G(Y)^{1/2} dY dy.$$

Then we can rewrite $I_N(\phi)$ as

$$I_N(\phi) = \sum_{T\in\mathcal{T}(G)} \mid W(G,T) \mid^{-1} \int_{\mathfrak{t}^0(F)} \int_{Z_G(F)\backslash G(F)} \hat{\phi}(g^{-1}Yg)\kappa_N(\gamma_Y^{-1}g) dg D^G(Y)^{1/2} dY.$$

For $T \in T(G)$ and $Y \in \mathfrak{t}^0(F)$, define the truncation function $\kappa_{N,T,Y}$ on $A_T(F)\backslash G(F)$ to be ($A_T(F)$ is the maximal split subtorus of $T(F)$)

$$\kappa_{N,T,Y}(g) = \int_{A_G(F)\backslash A_T(F)} \kappa_N(\gamma_Y^{-1}ag) da. \tag{4.5}$$

Then we have

(4.6)

$$I_N(\phi) = \sum_{T\in\mathcal{T}(G)} \mid W(G,T) \mid^{-1} \int_{\mathfrak{t}^0(F)} \int_{A_T(F)\backslash G(F)} \hat{\phi}(g^{-1}Yg)\kappa_{N,T,Y}(g) dg D^G(Y)^{1/2} dY.$$

The next step is to change the weight factor $\kappa_{N,T,Y}$ to the standard weight factor defined by Arthur. For $T \in \mathcal{T}(G)$, we fix a minimal parabolic subgroup $P_{min} = M_{min}N_{min}$ of G with $A_T(F) \subset M_{min}(F)$. For $Y_{min} \in \mathfrak{a}_{P_{min}}^{+}$, let $\tilde{v}_{T,Y_{min}}(g)$ be the weight factor defined by the same formula as equation (9.9) of [17], which is a function on $A_T(F)\backslash G(F)$. By the same argument as in Section 9.1 of [17], we can get rid of certain elements $Y \in \mathfrak{t}^0(F)$ that are not very regular with respect to N. To be specific, we only need to consider those $Y \in \mathfrak{t}^0(F)$ such that $|Q(Y)| > N^{-b}$ for all $Q \in \mathcal{Q}_T$ where

- $\mathcal{Q}_T$ is a finite set of polynomials on $\mathfrak{t}(F)$ defined in the same way as the Ginzburg-Rallis model case (Section 9.1 and 9.5 of [17]).
- $b > 1$ is a large positive integer so that the analogue of Lemma 9.2 of [17] holds for our case. This lemma allows us to get rid of the remaining $Y \in \mathfrak{t}^0(F)$.

For $Y \in \mathfrak{t}^0(F)$ such that $|Q(Y)| > N^{-b}$ for all $Q \in \mathcal{Q}_T$, we have the following proposition.

Proposition 4.8. *There exist $c > 0, b > 1$ and $N_0 > 1$ (all independent of Y) such that if $N \geq N_0$ and $c\log(N) < \inf\{\alpha(Y_{min})|\ \alpha \in \Delta(A_{M_{min}}, P_{min})\}$, we have*

$$\int_{A_T(F)\backslash G(F)} \hat{\phi}(g^{-1}Yg)\kappa_{N,T,Y}(g)dg = \int_{A_T(F)\backslash G(F)} \hat{\phi}(g^{-1}Yg)\tilde{v}_{T,Y_{min}}(g)dg.$$

Proof. This is the most technical step in the proof of the geometric expansion. However, due to the similarity between the model (G,H) and the Ginzburg-Rallis model, the proof is very similar to the Ginzburg-Rallis model case in Section 9 of [17]. To be specific, we can decompose both weight factors in terms of the parabolic subgroups of G (not necessarily proper, in particular there is a term corresponds to G) as in equation (9.12)-(9.19) of [17]. By the same argument as in statement (4) of the proof of Proposition 9.5 of [17], we can show that the terms correspond to G are equal to each other. Then by a similar argument as in statement (5)-(7) of the proof of Proposition 9.5 of [17], we can show that the terms correspond to proper parabolic subgroups are all equal to 0 by using the strongly cuspidal property of f. This proves the proposition. □

After changing the weight factors, by the same argument as in Section 9.7 and 10.1 of [17], we can show that

$$I(\phi) = \lim_{N\to\infty} I_N(\phi) = \sum_{T\in\mathcal{T}(G)} |W(G,T)|^{-1} \int_{\mathfrak{t}^0(F)} D^G(t)^{1/2}\hat{\theta}_{\phi}(t)\, \mathrm{d}t. \tag{4.7}$$

Lastly, we can apply the standard argument as in the GGP case (Section 11 of [2]) to finish the proof of the geometric side of the trace formula. To be specific, first, by a very similar semi-simple descent argument as in Section 11.6 of [2], we can reduce the proof of the geometric expansion to the case when f is supported on a small neighborhood of the center $Z_G(F)$. Note that in order to apply the argument in Section 11.6 of [2], we only need the following two inputs.

- An analogue of Proposition 11.5.1 of [2] for our model. This is just Proposition 4.4.
- The geometric side of the local trace formula holds for the model (G_x, H_x) where x is any element in $G(F)_{ell} \cap H_0(F)$ with $x \notin Z_G(F)$. For such x, the model (G_x, H_x) is just the Whittaker model of G_x, we can easily prove the geometric expansion for the model (G_x, H_x) by the same argument as in Section 10.4 of [17].

After we have reduced the proof of the geometric expansion to the case when f is supported on a small neighborhood of the center $Z_G(F)$, by the same argument as in Section 11.7 of [2], we only need to prove the analog of the geometric expansion on the Lie algebra level. The proof of the geometric expansion on Lie algebra follows from the same homogeneous argument in Sections 11.8-11.9 of [2] together with Lemma 4.6 and equation (4.7). This finishes the proof of the geometric side of the trace formula.

4.4. **The multiplicity formula.** Let π be a smooth finite length tempered representation of $G(F)$ with central character η. Define the geometric multiplicity $m_{geom}(\pi)$ to be

$$m_{geom}(\pi) := c_\pi(1) + \sum_{T \in \mathcal{T}_{ell}(H_0)} |W(H_0,T)|^{-1} \mathrm{vol}(T(F)/Z_G(F))^{-1} \int_{T(F)/Z_G(F)} D^H(t) c_\pi(t) \omega^{-1}(t) \, \mathrm{d}t.$$

Here $c_\pi(t) = c_{\theta_\pi}(t)$ is the regular germ of θ_π at t. Note that the expression of $m_{geom}(\pi)$ is almost the same as the geometric expansion $I_{geom}(f)$. The only difference is that we replace the quasi-character θ_f by θ_π.

Theorem 4.9. *The multiplicity formula*

$$m(\pi) = m_{geom}(\pi)$$

holds for all tempered representations.

In the next section, we will prove this multiplicity formula by assuming the trace formula holds. Apparently it is enough to prove it for irreducible tempered representations.

4.5. **The reduced models.** In this subsection, we will discuss the reduced models of the pair (G, H). With the notation as in Section 3, the reduced models are just the models $(G_{\bar{Q}}, H_{\bar{Q}})$ where $\bar{Q} = M_Q \bar{U}_Q$ runs over the good parabolic subgroups of G. This models will be used in the proof of the spectral side of the trace formula.

We first define the multiplicities for the reduced models. Let τ be a smooth finite length representation of $G_{\bar{Q}}(F)$ whose central character equals η on $Z_G(F)$. Define the multiplicity $m(\tau)$ to be

$$m(\tau) := \dim(Hom_{H_{\bar{Q}}(F)}(\tau, (\omega \otimes \xi)|_{H_{\bar{Q}}(F)} \otimes \delta_{H_{\bar{Q}}}^{1/2})).$$

Note that as in Proposition 3.5, when we consider the reduced models, we need to twist the extra modular character $\delta_{H_{\bar{Q}}}^{1/2}$. For simplicity, we will use $\omega_Q\otimes\xi_Q$ (or just ω_Q if ξ_Q is trivial) to denote the character $(\omega\otimes\xi)|_{H_{\bar{Q}}(F)}\otimes\delta_{H_{\bar{Q}}}^{1/2}$.

For our application, we need to divide the reduced models into two categories. We say the model $(G_{\bar{Q}},H_{\bar{Q}})$ is of *Type I* if it appears both in the quasi-split case and the non quasi-split case (this is the same situation as in the GR model case, see Appendix B of [17] for details). This is equivalent to say that the Levi subgroup $M_Q(F)$ is isomorphic to $\mathrm{GL}_1(E)\times\mathrm{GU}(J_{4,\varepsilon})(F)$, $\mathrm{GL}_2(E)\times\mathrm{GU}(J_{2,\varepsilon})(F)$ or $\mathrm{GL}_1(E)\times\mathrm{GL}_1(E)\times\mathrm{GU}(J_{2,\varepsilon})(F)$. The rest reduced models are called *Type II* reduced models. In particular, Type II reduced models only appear in the quasi-split case.

For the rest part of this subsection, we will describe the reduced models and the multiplicity formulas associated to them. **We first consider Type I reduced models.** If $M_Q(F)$ is isomorphic to $\mathrm{GL}_2(E)\times\mathrm{GU}(J_{2,\varepsilon})(F)$, the reduced model $(G_{\bar{Q}},H_{\bar{Q}})$ is just (G_0,H_0). And the character ω_Q on $H_{\bar{Q}}$ is just the character ω. In this case, the multiplicity formula is very similar to the (G,H) case: given a smooth finite length tempered representation τ of $G_0(F)$ whose central character equals η on $Z_{H_0}(F)$, define

$$m_{geom}(\tau):=c_\tau(1)+\sum_{T\in\mathcal{T}_{ell}(H_0)}|W(H_0,T)|^{-1}\mathrm{vol}(T(F)/Z_G(F))^{-1}\int_{T(F)/Z_G(F)}D^{H_0}(t)\theta_\tau(t)\omega^{-1}(t)\,\mathrm{d}t.$$

The multiplicity formula is just $m(\tau)=m_{geom}(\tau)$. The two models we get here are the only pure inner forms of each other. These two models are essentially the Gan–Gross–Prasad models for $\mathrm{SO}_4\times\mathrm{SO}_3$.

Remark 4.10. *Let G be a reductive group and H be a spherical subgroup of G, i.e. there exists a Borel subgroup B of G (not necessarily defined over F) such that BH is Zariski open in G. Following the definition in Section 16.5 of* [13]*, the pure inner forms of the spherical pair (G,H) are parameterized by the set $H^1(F,H/Z_{G,H})$ where $Z_{G,H}=Z_G\cap H$. More specifically, each element $\alpha\in H^1(F,H/Z_{G,H})$ induces an element in $H^1(F,G/Z_{G,H})$ (still denoted by α) via the map $H^1(F,H/Z_{G,H})\to H^1(F,G/Z_{G,H})$. This induces an inner form G_α (resp. H_α) of G (resp. H) with $H_\alpha\subset G_\alpha$. As a result, we get a spherical pair (G_α,H_α).*

If $M_Q(F)$ is isomorphic to $\mathrm{GL}_1(E)\times\mathrm{GU}(J_{4,\varepsilon})(F)$, the reduced model can be described as follows: $G_{\bar{Q}}(F)=M_Q(F)$, $H_{\bar{Q}}=H_{0,\bar{Q}}\ltimes U_{\bar{Q}}$ with

$$H_{0,\bar{Q}}(F)=\{h_Q(a,b)=\begin{pmatrix}a\end{pmatrix}\times diag(b,a,b,b)\colon a,b\in E^\times,\ N_{E/F}(a)=N_{E/F}(b)\},$$

$$U_{\bar{Q}}(F)=\{u_Q(x,y,z)=\begin{pmatrix}1\end{pmatrix}\times\begin{pmatrix}1&x&y&z\\0&1&0&\varepsilon^{-1}\bar{x}\\0&0&1&-\bar{y}\\0&0&0&1\end{pmatrix}:x,y,z\in E,\ z+\bar{z}-\varepsilon^{-1}x\bar{x}+y\bar{y}=0\}.$$

The character $\omega_Q \otimes \xi_Q$ on $H_{\bar{Q}}(F)$ is given by

$$\omega_Q \otimes \xi_Q(h_Q(a,b)u_Q(x,y,z)) = \chi_1(a)\chi_2(b)\psi(y+\bar{y})$$

where χ_1 and χ_2 are some unitary characters of $E^\times$ with $\eta = \chi_1\chi_2$. We define the geometric multiplicity to be

$$m_{geom}(\tau) := c_\tau(1)+\mathrm{vol}(H_{0,\bar{Q}}(F)/Z_G(F))^{-1}\int_{Z_G(F)\backslash H_{0,\bar{Q}}(F)} D^{H_{\bar{Q}}}(t)c_\tau(t)\omega_Q(t)\,\mathrm{d}t$$

where τ is a smooth finite length tempered representation of $G_{\bar{Q}}(F)$ whose central character equals η on $Z_G(F)$. The two models we get here are the only pure inner forms of each other. These two models are essentially the Gan–Gross–Prasad models for $\mathrm{U}_4 \times \mathrm{U}_1$.

If $M_Q(F)$ is isomorphic to $\mathrm{GL}_1(E) \times \mathrm{GL}_1(E) \times \mathrm{GU}(J_{2,\varepsilon})(F)$, the reduced model can be described as follows: $G_{\bar{Q}}(F) = M_Q(F)$, and

$$H_{\bar{Q}}(F) = H_{0,\bar{Q}}(F) = \{h_Q(a,b) = \big(a\big)\times\big(b\big)\times diag(a,b) \colon a,b \in E^\times,\ N_{E/F}(a) = N_{E/F}(b)\}.$$

The character ω_Q on $H_{\bar{Q}}$ is given by

$$\omega_Q(h_Q(a,b)) = \chi_1(a)\chi_2(b)$$

where χ_1 and χ_2 are some unitary characters of $E^\times$ with $\eta = \chi_1\chi_2$. We define the geometric multiplicity to be

$$m_{geom}(\tau) := c_\tau(1)+\mathrm{vol}(H_{0,\bar{Q}}(F)/Z_G(F))^{-1}\int_{Z_G(F)\backslash H_{0,\bar{Q}}(F)} D^{H_{\bar{Q}}}(t)c_\tau(t)\omega_Q(t)\,\mathrm{d}t$$

where τ is a smooth finite length tempered representation of $G_{\bar{Q}}(F)$ whose central character equals η on $Z_G(F)$. The two models we get here are the only pure inner forms of each other. These two models are essentially the Gan–Gross–Prasad models for $\mathrm{U}_2 \times \mathrm{U}_1$.

Remark 4.11. *By the description above, we know that all the Type I reduced models are essentially lower rank Gan–Gross–Prasad models. Also it is easy to see that the extra modular character $\delta_{H_{\bar{Q}}}^{1/2}$ is trivial for all Type I reduced models.*

Then we consider Type II reduced models. There are three Type II reduced models (all in the quasi-split case) which correspond to the cases when $M_Q(F)$ is isomorphic to $\mathrm{GL}_3(E)\times\mathrm{GL}_1(F)$, $\mathrm{GL}_2(E)\times\mathrm{GL}_1(E)\times\mathrm{GL}_1(F)$ and $\mathrm{GL}_1(E) \times \mathrm{GL}_1(E) \times \mathrm{GL}_1(E) \times \mathrm{GL}_1(F)$. When $M_Q(F) = \mathrm{GL}_3(E) \times \mathrm{GL}_1(F)$, up to modulo the $\mathrm{GL}_1(F)$-part which is abelian, the reduced model can be described as follows (i.e. $M_Q = G_{\bar{Q}} \times \mathrm{GL}_1$ and $H \cap \bar{Q} = H_{\bar{Q}} \times \mathrm{GL}_1$):

$G_{\bar{Q}}(F) = \mathrm{GL}_3(E)$ and $H_{\bar{Q}}(F) = H_{0,\bar{Q}}(F) \ltimes U_{\bar{Q}}(F)$ where

$$H_{0,\bar{Q}}(F) = \{\begin{pmatrix} a & 0 & 0 \\ c & b & 0 \\ 0 & 0 & a \end{pmatrix} : a, b \in E^{\times}, c \in E, \frac{a}{b} \in F^{\times}, \frac{c}{a} \in \sqrt{\alpha}F\},$$

$$U_{\bar{Q}}(F) = \{\begin{pmatrix} 1 & 0 & x_1 \\ 0 & 1 & x_2 \\ 0 & 0 & 1 \end{pmatrix} : x_1,\ x_2 \in E\}.$$

And the character $\omega_Q \otimes \xi_Q$ is given by

$$\omega_Q \otimes \xi_Q(\begin{pmatrix} a & 0 & 0 \\ c & b & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} 1 & 0 & x_1 \\ 0 & 1 & x_2 \\ 0 & 0 & 1 \end{pmatrix}) = |\frac{a}{b}|^{-3/2} \chi_1(a)\chi_2(b)\psi(x_1 + \bar{x}_1)$$

where χ_1 and χ_2 are some unitary characters of $E^{\times}$ with $\eta = \chi_1\chi_2$. Here the factor $|\frac{a}{b}|^{-3/2}$ comes from the extra modular character $\delta_{H_{\bar{Q}}}^{1/2}$.

When $M_Q = \mathrm{GL}_2(E) \times \mathrm{GL}_1(E) \times \mathrm{GL}_1(F)$, up to modulo the $\mathrm{GL}_1(F)$ and $\mathrm{GL}_1(E)$ parts which are abelian, the reduced model can be described as follows: $G_{\bar{Q}}(F) = \mathrm{GL}_2(E)$, and

$$H_{\bar{Q}}(F) = H_{0,\bar{Q}}(F) = \{\begin{pmatrix} a & 0 \\ c & b \end{pmatrix} : a, b \in E^{\times}, c \in E, \frac{a}{b} \in F^{\times}, \frac{c}{a} \in \sqrt{\alpha}F\}.$$

And the character ω_Q is given by

$$\omega_Q(\begin{pmatrix} a & 0 \\ c & b \end{pmatrix}) = |\frac{a}{b}|^{-1/2} \chi_1(a)\chi_2(b)$$

where χ_1 and χ_2 are some unitary characters of $E^{\times}$ with $\eta = \chi_1\chi_2$, and the factor $|\frac{a}{b}|^{-1/2}$ comes from the extra modular character $\delta_{H_{\bar{Q}}}^{1/2}$.

When $M_Q = \mathrm{GL}_1(E) \times \mathrm{GL}_1(E) \times \mathrm{GL}_1(E) \times \mathrm{GL}_1(F)$, the model is abelian and the multiplicity is trivially equal to 1 for all irreducible representations (which are just characters).

For all the Type II reduced models, the geometric multiplicity is defined to be (again τ is a smooth finite length tempered representation of $G_{\bar{Q}}(F)$ whose central character equals η on $Z_G(F)$)

$$m_{geom}(\tau) = c_{\tau}(1).$$

Moreover, it is easy to see that $G_{\bar{Q}}$ is a product of the general linear groups and $H^1(H_{\bar{Q}}/Z_{G_{\bar{Q}}, H_{\bar{Q}}}, F)$ is trivial for all Type II reduced models. Hence all these models don't have any other pure inner form and each L-packet only contains one element.

Remark 4.12. *The most important feature of the Type II reduced models is that there is no elliptic element in $H_{\bar{Q}}(F)$ other than the center. As a result, in the multiplicity formulas for these models, we only have the germ at the identity element. This is an analogue of the Type II reduced models for the GR model (see Appendix B of* [17]*).*

The proof of the multiplicity formula for the reduced models follows from the same, but easier arguments as the proof of the multiplicity formula for the model (G,H). To be specific, if $\bar{Q}$ is the Borel subgroup, the multiplicity formula is trivial. If $\bar{Q}$ is not the Borel subgroup, by induction we may assume that the multiplicity formula holds for all reduced models $(G_{\bar{Q}'}, H_{\bar{Q}'})$ with $\bar{Q}' \subsetneq \bar{Q}$. Note that the models $\{(G_{\bar{Q}'}, H_{\bar{Q}'})|\ \bar{Q}' \subsetneq \bar{Q}\}$ are the reduced models of $(G_{\bar{Q}}, H_{\bar{Q}})$. By the same argument as in the proof of the multiplicity formula for the model (G,H) in the next section, we can then prove the multiplicity formula for the model $(G_{\bar{Q}}, H_{\bar{Q}})$ (the proof in the next section uses the fact that the multiplicity formula holds for all the reduced models of (G,H)). Hence from now on, we will assume that the multiplicity formulas hold for all the reduced models.

4.6. **Some consequences of the multiplicity formulas for the reduced models.** Let $(G_{\bar{Q}}, H_{\bar{Q}})$ be a reduced model and let Π be a tempered local Vogan L-packet of $G_{\bar{Q}}(F)$ whose central character equals η on $Z_G(F)$. If the reduced model is of Type II, $G_{\bar{Q}}$ is the product of some general linear groups, there is no other pure inner form of the model $(G_{\bar{Q}}, H_{\bar{Q}})$ and the L-packet Π only contains one element τ. If the reduced model is of Type I, then there is another pure inner form of the model $(G_{\bar{Q}}, H_{\bar{Q}})$ (as we described in the previous subsection) and the L-packet Π contains representations of both groups and may have more than one element.

Theorem 4.13. *(1)* $\sum_{\tau\in\Pi} m(\tau)=1.$

(2) $m(\tau)\leq 1$ *for all irreducible tempered representations* τ *of* $G_{\bar{Q}}(F)$ *whose central character equals* η *on* $Z_G(F)$.

Proof. (2) is a direct consequence of (1). For (1), if the reduced model is of Type II, by the discussion above, the L-packet Π only contains one element τ which is an irreducible tempered representation of some general linear groups. In particular, τ is a generic representation. Combining the multiplicity formula in the previous subsection and the work of Rodier in [11], we have

$$\sum_{\tau\in\Pi} m(\tau)=m(\tau)=m_{geom}(\tau)=c_{\tau}(1)=1.$$

This proves (1) for Type II reduced models.

Next, we consider the Type I reduced models. We will only consider the case when $G_{\bar{Q}}(F)\simeq \mathrm{GL}_2(E)\times \mathrm{GU}(J_{2,\varepsilon})(F)$. The arguments for the rest two cases are similar. For $i=1,2$, fix $\varepsilon_i\in F^{\times}$ with $\eta_{E/F}(\varepsilon_i)=(-1)^{i-1}$ and let $G_{\bar{Q},\varepsilon_i}(F)=\mathrm{GL}_2(E)\times \mathrm{GU}(J_{2,\varepsilon_i})(F)$. Then the L-packet Π is of the form

$$\Pi=\{\tau_0\otimes\tau_{\varepsilon_1}:\tau_{\varepsilon_1}\in\Pi_{\phi}(\mathrm{GU}(J_{2,\varepsilon_1}))\}\cup\{\tau_0\otimes\tau_{\varepsilon_2}:\tau_{\varepsilon_2}\in\Pi_{\phi}(\mathrm{GU}(J_{2,\varepsilon_2}))\}$$

where τ_0 is an irreducible tempered representation of $\mathrm{GL}_2(E)$, and $\Pi_{\phi}=\Pi_{\phi}(\mathrm{GU}(J_{2,\varepsilon_1}))\cup\Pi_{\phi}(\mathrm{GU}(J_{2,\varepsilon_2}))$ is a tempered local L-packet of $\mathrm{GU}_2(F)$. By the multiplicity formula in the previous subsection, we know that $\sum_{\tau\in\Pi} m(\tau)$ is equal to

$$\sum_{i=1}^{2} \sum_{\tau_{\varepsilon_i}\in\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_i}))} c_{\tau_0}(1)c_{\tau_{\varepsilon_i}}(1) + \sum_{i=1}^{2} \sum_{\tau_{\varepsilon_i}\in\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_i}))} \sum_{T_i\in\mathcal{T}_{ell}(\mathrm{GU}(J_{2,\varepsilon_i}))} \times \nu(T_i)\int_{T_i(F)/Z_{\mathrm{GU}(J_{2,\varepsilon_i})}(F)} D^{\mathrm{GU}(J_{2,\varepsilon_i})}(t_i)\theta_{\tau_0}(t_i)\theta_{\tau_{\varepsilon_i}}(t_i)\omega_{\varepsilon_i}^{-1}(t_i)\,\mathrm{d}t_i$$

with $\nu(T_i) = |W(\mathrm{GU}(J_{2,\varepsilon_i}),T_i)|^{-1}\mathrm{vol}(T_i(F)/Z_{\mathrm{GU}(J_{2,\varepsilon_i})}(F))^{-1}$. Since τ_0 is a tempered representation of $\mathrm{GL}_2(E)$, it is generic and hence $c_{\tau_0}(1) = 1$ by the work of Rodier in [11]. Combining with Conjecture 2.5(2), we have

$$\sum_{i=1}^{2} \sum_{\tau_{\varepsilon_i}\in\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_i}))} c_{\tau_0}(1)c_{\tau_{\varepsilon_i}}(1) = \sum_{i=1}^{2} \sum_{\tau_{\varepsilon_i}\in\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_i}))} c_{\tau_{\varepsilon_i}}(1) = 1.$$

Hence $\sum_{\tau\in\Pi} m(\tau)$ equals

$$1+\sum_{i=1}^{2} \sum_{T_i\in\mathcal{T}_{ell}(\mathrm{GU}(J_{2,\varepsilon_i}))} \nu(T_i)\int_{T_i(F)/Z_{\mathrm{GU}(J_{2,\varepsilon_i})}(F)} D^{\mathrm{GU}(J_{2,\varepsilon_i})}(t_i)\theta_{\tau_0}(t_i)\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_i}))}(t_i)\omega_{\varepsilon_i}^{-1}(t_i)\,\mathrm{d}t_i.$$

Here we recall from Conjecture 2.5 that $\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_i}))} = \sum_{\tau_{\varepsilon_i}\in\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_i}))} \theta_{\tau_{\varepsilon_i}}$.

Now we are ready to prove the theorem. We have a natural bijection

$$T_1 \in \mathcal{T}_{ell}(\mathrm{GU}(J_{2,\varepsilon_1})) \leftrightarrow T_2 \in \mathcal{T}_{ell}(\mathrm{GU}(J_{2,\varepsilon_2}))$$

between the conjugacy classes of maximal elliptic tori of $\mathrm{GU}(J_{2,\varepsilon_1})(F)$ and $\mathrm{GU}(J_{2,\varepsilon_2})(F)$. Hence in order to prove the theorem, it is enough to show that for all $T_1 \leftrightarrow T_2$, we have

$$\nu(T_1)\int_{T_1(F)/Z_{\mathrm{GU}(J_{2,\varepsilon_1})}(F)} D^{\mathrm{GU}(J_{2,\varepsilon_1})}(t_1)\theta_{\tau_0}(t_1)\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_1}))}(t_1)\omega_{\varepsilon_1}^{-1}(t_1)\,\mathrm{d}t_1$$

$$= -\,\nu(T_2)\int_{T_2(F)/Z_{\mathrm{GU}(J_{2,\varepsilon_2})}(F)} D^{\mathrm{GU}(J_{2,\varepsilon_2})}(t_2)\theta_{\tau_0}(t_2)\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_2}))}(t_2)\omega_{\varepsilon_2}^{-1}(t_2)\,\mathrm{d}t_2.$$

We fix such a pair (T_1,T_2). It is easy to see that $\nu(T_1)=\nu(T_2)$. For $t_1 \in T_1(F)$ and $t_2 \in T_2(F)$, we write $t_1 \leftrightarrow t_2$ if they have the same characteristic polynomial. Then it is enough to show that for all $t_1 \in T_1(F)_{reg}$ and $t_2 \in T_2(F)_{reg}$ with $t_1 \leftrightarrow t_2$, we have

$$D^{\mathrm{GU}(J_{2,\varepsilon_1})}(t_1)\theta_{\tau_0}(t_1)\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_1}))}(t_1)\omega_{\varepsilon_1}^{-1}(t_1) = -D^{\mathrm{GU}(J_{2,\varepsilon_2})}(t_2)\theta_{\tau_0}(t_2)\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_2}))}(t_2)\omega_{\varepsilon_2}^{-1}(t_2). \tag{4.8}$$

Since $t_1 \leftrightarrow t_2$, we have

$$D^{\mathrm{GU}(J_{2,\varepsilon_1})}(t_1)\theta_{\tau_0}(t_1)\omega_{\varepsilon_1}^{-1}(t_1) = D^{\mathrm{GU}(J_{2,\varepsilon_2})}(t_2)\theta_{\tau_0}(t_2)\omega_{\varepsilon_2}^{-1}(t_2).$$

By Conjecture 2.5(3), we also have $\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_1}))}(t_1) = -\theta_{\Pi_\phi(\mathrm{GU}(J_{2,\varepsilon_2}))}(t_2)$ (note that Conjecture 2.5 is known for GU_2 by the work of Jacquet–Langlands [7]). This proves (4.8) and completes the proof of the theorem. □

The following proposition will be proved in Appendix A by Mackey theory.

Proposition 4.14. *Let $\bar{Q} = M_Q\bar{U}_Q$ be a good parabolic subgroup of G, and let τ be a smooth finite length tempered representation of $M_Q(F)$ whose central character equals η on $Z_G(F)$. Set $\pi = I_{\bar{Q}}^G(\tau)$. Then*

$$m(\pi) \leq m(\tau).$$

Corollary 4.15. *For all $\pi \in \Pi_{temp}(G,\eta) \smallsetminus \Pi_2(G,\eta)$, we have*

$$m(\pi) \leq 1.$$

Proof. For $\pi \in \Pi_{temp}(G,\eta) \smallsetminus \Pi_2(G,\eta)$, there exists a good proper parabolic subgroup $\bar{Q} = M_Q\bar{N}_Q$ and an irreducible tempered representation τ of $M_Q(F)$ such that $\pi = I_{\bar{Q}}^G(\tau)$. By Theorem 4.13 (2) and Proposition 4.14, we have

$$m(\pi) = m(I_{\bar{Q}}^G(\tau)) \leq m(\tau) \leq 1.$$

This proves the corollary. □

The following theorem is a stronger version of Proposition 4.14. It will be proved in Section 6.1 by assuming Proposition 4.14 holds.

Theorem 4.16. *With the same assumptions and notation as in Proposition 4.14, we have $m(\pi) = m(\tau)$.*

5. The proof of the main theorems

In this section, we are going to prove our main theorems by assuming the trace formula in Theorem 4.3 holds. We will prove the unitary similitude group case in Section 5.1, and the unitary group case in Section 5.2.

5.1. **The unitary similitude group case.** We first prove the multiplicity formula $m(\pi) = m_{geom}(\pi)$ for all tempered representations. For simplicity, we still use (G,H) to denote $(G_\varepsilon, H_\varepsilon)$ and $\omega \otimes \xi$ to denote $\omega_\varepsilon \otimes \xi_\varepsilon$. We need a proposition.

Proposition 5.1. *Let $\bar{Q} = M_Q\bar{U}_Q$ be a good parabolic subgroup of G, and let τ be an irreducible tempered representation $M_Q(F)$ whose central character equals η on $Z_G(F)$. Set $\pi = I_{\bar{Q}}^G(\tau)$. Then*

$$m(\pi) = m(\tau), \;\; m_{geom}(\pi) = m_{geom}(\tau). \tag{5.1}$$

Proof. The first equality follows from Theorem 4.16, while the second equality is a direct consequence of Lemma 2.3 of [16] together with the definitions of $m_{geom}(\pi)$ and $m_{geom}(\tau)$. □

Now the multiplicity formula will be a direct consequence of the trace formula in Theorem 4.3, together with (5.1). The argument is the same as the GGP case (Proposition 11.5.1 of [2]), we will only give a sketch of the proof. First, as in Proposition 11.5.1(ii) of [2], using the trace formula, (5.1), and the assumption that the multiplicity formulas hold for all reduced models, we know that for all strongly cuspidal function $f \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F), \eta^{-1})$, we have

$$\sum_{\pi \in \mathcal{X}_{ell}(G,\eta)} D(\pi)\theta_f(\pi)(m(\bar{\pi}) - m_{geom}(\bar{\pi})) = 0. \tag{5.2}$$

Then by Corollary 5.7.2(iv) of [2], for each $\pi \in \mathcal{X}_{ell}(G,\eta)$, there exists $f_\pi \in \mathcal{C}_{scusp}(Z_G(F)\backslash G(F), \eta^{-1})$ such that for all $\pi' \in \mathcal{X}_{ell}(G,\eta)$, $\theta_f(\pi') \neq 0$ if and only if $\pi = \pi'$. Put f_π in the equation (5.2), we have

$$D(\pi)\theta_{f_\pi}(\pi)(m(\bar{\pi}) - m_{geom}(\bar{\pi})) = 0,$$

which implies that $m(\pi) = m_{geom}(\pi)$. This proves the multiplicity formula.

Now we are ready to prove Theorem 1.1. For $i = 1, 2$, fix $\varepsilon_i \in F^\times$ with $\eta_{E/F}(\varepsilon_i) = (-1)^{i-1}$ as before. Let Π_ϕ be a local tempered Vogan L-packet of GU_6. Then we can write Π_ϕ as $\Pi_\phi = \Pi_\phi(G_{\varepsilon_1}) \cup \Pi_\phi(G_{\varepsilon_2})$. Our goal is to show that

$$\sum_{\pi \in \Pi_\phi} m(\pi) = \sum_{i=1}^{2} \sum_{\pi_{\varepsilon_i} \in \Pi_\phi(G_{\varepsilon_i})} m(\pi_{\varepsilon_i}) = 1. \tag{5.3}$$

By the multiplicity formula, we have

$$\sum_{\pi \in \Pi_\phi} m(\pi) = \sum_{i=1}^{2} \sum_{\pi_{\varepsilon_i} \in \Pi_\phi(G_{\varepsilon_i})} \left(c_{\pi_{\varepsilon_i}}(1) + \sum_{T_i \in \mathcal{T}_{ell}(H_{0,\varepsilon_i})} \nu(T_i) \int_{T_i(F)/Z_{G_{\varepsilon_i}}(F)} D^{H_{\varepsilon_i}}(t_i) c_{\pi_{\varepsilon_i}}(t_i) \omega_{\varepsilon_i}^{-1}(t_i)\, \mathrm{d}t_i \right)$$

with $\nu(T_i) = |W(H_{0,\varepsilon_i}, T_i)|^{-1} \mathrm{vol}(T_i(F)/Z_{G_{\varepsilon_i}}(F))^{-1}$. By the same argument as in the proof of Theorem 4.13, together with Conjecture 2.5 (2) and (3), we have

$$\sum_{i=1}^{2} \sum_{\pi_{\varepsilon_i} \in \Pi_\phi(G_{\varepsilon_i})} c_{\pi_{\varepsilon_i}}(1) = 1,$$

$$\sum_{i=1}^{2} \sum_{\pi_{\varepsilon_i} \in \Pi_\phi(G_{\varepsilon_i})} \sum_{T_i \in \mathcal{T}_{ell}(H_{0,\varepsilon_i})} \nu(T_i) \int_{T_i(F)/Z_{G_{\varepsilon_i}}(F)} D^{H_{\varepsilon_i}}(t_i) c_{\pi_{\varepsilon_i}}(t_i) \omega_{\varepsilon_i}^{-1}(t_i)\, \mathrm{d}t_i = 0.$$

This proves (5.3) and finishes the proof of Theorem 1.1.

5.2. **The unitary group case.** In this subsection, we are going to prove Theorem 1.2. The idea is to study the relations between the models associated to unitary similitude groups and the models associated to unitary groups. We first recall the definition of the character ω_ε (resp. $\omega_{1,\varepsilon}$) of $H_{0,\varepsilon}(F)$ (resp. $H_{0,1,\varepsilon}(F)$) in Section 1. We recall

$$\omega_\varepsilon(m(h,h)) = \chi_E(\det(h))\chi_F(\lambda(h)), \;\; \omega_{1,\varepsilon}(m(h_1,h_1)) = \chi_E(\det(h_1))$$

for $h \in \mathrm{GU}(J_{2,\varepsilon})(F)$ and $h_1 \in \mathrm{U}(J_{2,\varepsilon})(F)$ where χ_E (resp. χ_F) is a character of $E^\times$ (resp. $F^\times$), λ is the similitude character, and

$$m(h,h) = diag(h, h, \lambda(h) w_2 {}^t\bar{h}^{-1} w_2) \in H_{0,\varepsilon}(F), \;\; m(h_1,h_1) = diag(h_1, h_1, w_2 {}^t\bar{h}_1^{-1} w_2) \in H_{0,1,\varepsilon}(F).$$

Let $\eta_{E/F} : F^\times \to \mathbb{C}^\times$ be the quadratic character associated to E as before. We define a new character ω'_ε of $H_{0,\varepsilon}(F)$ to be

$$\omega'_\varepsilon(m(h,h)) = \chi_E(\det(h))\chi_F(\lambda(h))\eta_{E/F}(\lambda(h)),\ h \in \mathrm{GU}(J_{2,\varepsilon})(F).$$

And for any smooth finite length representation π_ε of $G_\varepsilon(F)$ with central character $\eta = \chi_E^2 \otimes (\chi_F \circ N_{E/F})$ (note that ω_ε is equal to ω'_ε on the center of $G_\varepsilon(F)$), we define the multiplicity

$$m(\pi_\varepsilon)' = \dim(\mathrm{Hom}_{H_\varepsilon(F)}(\pi_\varepsilon, \omega'_\varepsilon \otimes \xi_\varepsilon)).$$

The next proposition is essential in the proof of Theorem 1.2.

Proposition 5.2. *For any irreducible smooth representation π_ε of $G_\varepsilon(F) = \mathrm{GU}(J_{6,\varepsilon})(F)$ with central character η, let $\pi_{1,\varepsilon}$ be the restriction of π_ε to $G_{1,\varepsilon}(F) = \mathrm{U}(J_{6,\varepsilon})(F)$ which is a smooth finite length representation of $G_{1,\varepsilon}(F)$ with central character $\eta_1 = \chi_E^2|_{E^1} = \eta|_{E^1}$. Then*

$$m(\pi_{1,\varepsilon}) = m(\pi_\varepsilon) + m(\pi_\varepsilon)'.$$

Proof. We extend the character $\omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon}$ to $H_{1,\varepsilon}(F)Z_{H_\varepsilon}(F)$ by making it equal to $\omega_\varepsilon \otimes \xi_\varepsilon$ on $Z_{H_\varepsilon}(F)$, we still denote this character by $\omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon}$. Then we have

$$\mathrm{Hom}_{H_{1,\varepsilon}(F)}(\pi_{1,\varepsilon}, \omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon}) \simeq \mathrm{Hom}_{H_{1,\varepsilon}(F)Z_{H_\varepsilon}(F)}(\pi_\varepsilon|_{H_{1,\varepsilon}(F)Z_{H_\varepsilon}(F)}, \omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon}).$$

By Frobenius reciprocity, we have

$$\mathrm{Hom}_{H_{1,\varepsilon}(F)Z_{H_\varepsilon}(F)}(\pi_\varepsilon|_{H_{1,\varepsilon}(F)Z_{H_\varepsilon}(F)}, \omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon}) \simeq \mathrm{Hom}_{H_\varepsilon(F)}(\pi_\varepsilon, Ind_{H_{1,\varepsilon}(F)Z_{H_\varepsilon}(F)}^{H_\varepsilon(F)}(\omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon})).$$

Then the proposition follows from the fact that

$$Ind_{H_{1,\varepsilon}(F)Z_{H_\varepsilon}(F)}^{H_\varepsilon(F)}(\omega_{1,\varepsilon} \otimes \xi_{1,\varepsilon}) \simeq \omega_\varepsilon \otimes \xi_\varepsilon \oplus \omega'_\varepsilon \otimes \xi_\varepsilon.$$

□

Then we prove a multiplicity formula for the model $(G_{1,\varepsilon}, H_{1,\varepsilon})$. To simplify the notation, we will omit the subscript ε. We start with a lemma.

Lemma 5.3. *Let π be an irreducible tempered representation of $G(F)$ with central character η, and let π_1 be the restriction of π to $G_1(F)$. Then we have $m(\pi_1) = m_{geom}(\pi_1)$ where the geometric multiplicity $m_{geom}(\pi_1)$ is defined to be*

$$m_{geom}(\pi_1) := 2c_{\pi_1}(1) + \sum_{T_1 \in \mathcal{T}_{ell}(H_{0,1})} |W(H_{0,1}, T_1)|^{-1} \int_{T_1(F)/Z_{G_1}(F)} D^{H_1}(t_1) c_{\pi_1}(t_1) \omega_1(t_1)^{-1} \, \mathrm{d}_v t_1.$$

Here $c_{\pi_1}(t) = c_{\theta_{\pi_1}}(t)$ is the regular germ of θ_{π_1} at t and $\mathrm{d}_v t_1$ is the Haar measure on $T_1(F)$ such that the volume of $T_1(F)/Z_{G_1}(F)$ is equal to 1.

Proof. Combining Proposition 5.2 with the multiplicity formula for the unitary similitude group case, we have

(5.4)

$$m(\pi_1) = 2c_\pi(1) + \sum_{T \in \mathcal{T}_{ell}(H_0)} |W(H_0, T)|^{-1} \int_{T(F)/Z_G(F)} D^H(t) c_\pi(t) \omega(t)^{-1} (1 + \eta_{E/F}(\lambda(t))) \, \mathrm{d}_v t$$

where $\mathrm{d}_v t$ is the Haar measure on $T(F)$ such that the volume of $T(F)/Z_G(F)$ is equal to 1. For each $T \in \mathcal{T}_{ell}(H_0)$, let $T_1 = T \cap H_{0,1}$ which is a maximal elliptic torus of $H_{0,1}$. There is a bijection between the set $\mathcal{T}_{ell}(H_0)$ and the set of quadratic extensions of F. If T corresponds to a quadratic extension other than E, there exist $\gamma \in T(F)$ such that $T(F) = T_1(F)Z_G(F) \cup \gamma T_1(F)Z_G(F)$ and $\ker(\eta_{E/F} \circ \lambda|_T) = T_1(F)Z_G(F)$. This implies that

$$\int_{T(F)/Z_G(F)} D^H(t)c_\pi(t)\omega(t)^{-1}(1+\eta_{E/F}(\lambda(t)))\,\mathrm{d}_v t =$$

$$2\int_{T(F)/Z_G(F)} D^H(t)c_\pi(t)\omega(t)^{-1}1_{T_1(F)Z_G(F)}(t)\,\mathrm{d}_v t = \int_{T_1(F)/Z_{G_1}(F)} D^H(t_1)c_\pi(t_1)\omega(t_1)^{-1}\,\mathrm{d}_v t_1$$

where $\mathrm{d}_v t_1$ is the Haar measure on $T_1(F)$ such that the volume of $T_1(F)/Z_{G_1}(F)$ is equal to 1. If T corresponds to E, then $\eta_{E/F} \circ \lambda$ is trivial on $T(F)$ and $T(F) = T_1(F)Z_G(F)$. We have

$$\int_{T(F)/Z_G(F)} D^H(t)c_\pi(t)\omega(t)^{-1}(1+\eta_{E/F}(\lambda(t)))\,\mathrm{d}_v t = 2\int_{T_1(F)/Z_{G_1(F)}} D^H(t_1)c_\pi(t_1)\omega(t_1)^{-1}\,\mathrm{d}_v t_1$$

where $\mathrm{d}_v t_1$ is the Haar measure on $T_1(F)$ such that the volume of $T_1(F)/Z_{G_1}(F)$ is equal to 1. Combining the above two equations with (5.4), we have

$$m(\pi_1) = 2c_\pi(1) + \sum_{T \in \mathcal{T}_{ell}(H_0)} |W(H_0,T)|^{-1}\mu(T)\int_{T_1(F)/Z_{G_1}(F)} D^H(t_1)c_\pi(t_1)\omega(t_1)^{-1}\,\mathrm{d}_v t_1$$

where $\mu(T)$ is equal to 2 if T corresponds to E, and is equal to 1 otherwise.

Since π_1 is the restriction of π to $G_1(F)$, we have

$$c_\pi(1) = c_{\pi_1}(1) \text{ and } c_\pi(t_1) = c_{\pi_1}(t_1), \text{ for all } t_1 \in T_1(F)_{reg}.$$

Moreover it is easy to see from the definition that

$$D^H(t_1) = D^{H_1}(t_1), \omega(t_1) = \omega_1(t_1),\ \forall t_1 \in T_1(F).$$

Hence we have

(5.5)

$$m(\pi_1) = 2c_{\pi_1}(1) + \sum_{T \in \mathcal{T}_{ell}(H_0)} |W(H_0,T)|^{-1}\mu(T)\int_{T_1(F)/Z_{G_1}(F)} D^{H_1}(t_1)c_{\pi_1}(t_1)\omega_1(t_1)^{-1}\,\mathrm{d}_v t_1.$$

To end the proof, we only need to describe the relations between $\mathcal{T}_{ell}(H_0)$ and $\mathcal{T}_{ell}(H_{0,1})$. There are two cases. When $\eta_{E/F}(-\varepsilon) = 1$, we have a surjective map from $\mathcal{T}_{ell}(H_{0,1})$ to the set of all the quadratic extensions of F. The fiber of this map has two elements at E, and has one element at all the other quadratic extensions. If $T_1' \in \mathcal{T}_{ell}(H_{0,1})$ maps to a quadratic extension other than E, let $T \in \mathcal{T}_{ell}(H_0)$ corresponds to the same quadratic extension. Then $T_1 = T \cap H_{0,1}$ is a maximal elliptic torus of $H_{0,1}$ which is conjugated to T_1' and we also have $|W(H_0,T)| = |W(H_{0,1},T_1)|$. On the other hand, if $T_1', T_1'' \in \mathcal{T}_{ell}(H_{0,1})$ are the fiber of E, we can choose two maximal elliptic tori T and T_0 of H_0 such that both T and T_0 correspond to E and $T_1 = T \cap H_{0,1}$ (resp. $T_{1,0} = T_0 \cap H_{0,1}$) is a maximal elliptic torus of $H_{0,1}$

which is conjugated to T_1' (resp. T_1''). Moreover, in this case, we also have $|W(H_0,T)| = |W(H_{0,1},T_1)|$ and $|W(H_0,T_0)| = |W(H_{0,1},T_{1,0})|$. The upshot is that T and T_0 are conjugated to each other in H_0, but not in $H_{0,1}$.

If $\eta_{E/F}(-\varepsilon) = -1$, we have a bijection map from $\mathcal{T}_{ell}(H_{0,1})$ to the set of all the quadratic extensions of F. If $T_1' \in \mathcal{T}_{ell}(H_{0,1})$ maps to a quadratic extension other than E, let $T \in \mathcal{T}_{ell}(H_0)$ corresponds to the same quadratic extension. Then $T_1 = T \cap H_{0,1}$ is a maximal elliptic torus of $H_{0,1}$ which is conjugated to T_1' and we also have $|W(H_0,T)| = |W(H_{0,1},T_1)|$. On the other hand, if $T_1' \in \mathcal{T}_{ell}(H_{0,1})$ maps to the quadratic extension E, let $T \in \mathcal{T}_{ell}(H_0)$ corresponds to the quadratic extension E. Then $T_1 = T \cap H_{0,1}$ is a maximal elliptic torus of $H_{0,1}$ which is conjugated to T_1'. Moreover, in this case, we have $|W(H_0,T)| = 2$ and $|W(H_{0,1},T_1)| = 1$. Now the lemma will follow from (5.5) and the discussion above. □

Remark 5.4. *Since $c_\pi(t_1) = c_{\pi_1}(t_1)$ for $t_1 \in T_1(F)_{reg}$, we know that the integral defining $m_{geom}(\pi_1)$ is absolutely convergent if π_1 is the restriction of a tempered representation of $G(F)$ to $G_1(F)$ (see Definition 4.1). However, for a general tempered representation π_1 of $G_1(F)$, the integral defining $m_{geom}(\pi_1)$ is not necessarily convergent because the Lie algebra of $G_1(F)$ has more than one regular nilpotent orbit. In this case, following Section 11.2 of* [2]*, we define*

$$m_{geom}(\pi_1) = \lim_{s\to 0^+} m_{geom,s}(\pi_1)$$

where

$$m_{geom,s}(\pi_1) = 2c_{\pi_1}(1) + \sum_{T_1\in\mathcal{T}_{ell}(H_{0,1})} |W(H_{0,1},T_1)|^{-1} \int_{T_1(F)/Z_{G_1}(F)} D^{G_1}(t_1)^{1/2}$$
$$\Delta(t_1)^{s-1/2} c_{\pi_1}(t_1)\omega_1(t_1)^{-1}\, \mathrm{d}_v t_1,\ \Delta(t_1) = D^{G_1}(t_1)D^{H_1}(t_1)^{-2}.$$

By the same argument as in Proposition 11.2.1 of [2]*, we can show that the integral defining $m_{geom,s}(\pi_1)$ is absolutely convergent when $Re(s) > 0$ and the limit exists. Moreover, if the integral is convergent when $s = 0$ (for example, this is the case when π_1 is the restriction of a tempered representation of $G(F)$ to $G_1(F)$), then we have*

$$m_{geom}(\pi_1) = \lim_{s\to 0^+} m_{geom,s}(\pi_1) = m_{geom,0}(\pi_1).$$

Proposition 5.5. *Let π_1 be an irreducible tempered representation of $G_1(F)$ with central character η_1. Then $m(\pi_1) = m_{geom}(\pi_1)$.*

Proof. By Lemmas 2.3 and 2.4, there exists an irreducible tempered representation π of $G(F)$ with central character η such that π_1 is a direct summand of $\pi|_{G_1}$. If $\pi_1 = \pi|_{G_1}$, then the multiplicity formula has been proved in Lemma 5.3. If not, by Lemma 2.2, we have $\pi|_{G_1} = \pi_1 \oplus \pi_2$ with $\pi_2 = \pi_1 \circ Ad(g)$ is another irreducible tempered representation of $G_1(F)$ with central character η_1. Here g is an element in $G(F)$ with $\lambda(g) \notin \mathrm{Im}(N_{E/F})$. It is easy to see that we may choose $g \in H_0(F)$. By Lemma 5.3, we have

$$m(\pi_1) + m(\pi_2) = m(\pi|_{G_1}) = m_{geom}(\pi|_{G_1}) = m_{geom}(\pi_1) + m_{geom}(\pi_2).$$

Moreover, it is easy to see that the map

$$\mathrm{Hom}_{H_1(F)}(\pi_1,\omega_1\otimes\xi_1)\to\mathrm{Hom}_{H_1(F)}(\pi_2,\omega_1\otimes\xi_1):\quad l\mapsto l\circ\pi(g)$$

is an isomorphism. This implies that $m(\pi_1)=m(\pi_2)$. Hence in order to prove the multiplicity formula, it is enough to show that

$$m_{geom}(\pi_1)=m_{geom}(\pi_2). \tag{5.6}$$

Since $\pi_2=\pi_1\circ Ad(g)$, we have

$$c_{\pi_1}(1)=c_{\pi_2}(1). \tag{5.7}$$

On the other hand, $Ad(g)$ induces an automorphism of $H_{0,1}(F)$ and also induces a permutation of the conjugacy classes of maximal elliptic tori that preserves the associated Weyl groups, as a result, we have

$$\sum_{T_1\in\mathcal{T}_{ell}(H_{0,1})}|W(H_{0,1},T_1)|^{-1}\int_{T_1(F)/Z_{G_1}(F)}D^{G_1}(t_1)^{1/2}\Delta(t_1)^{s-1/2}c_{\pi_1}(t_1)\omega_1(t_1)^{-1}\,\mathrm{d}_vt_1$$

$$=\sum_{T_1\in\mathcal{T}_{ell}(H_{0,1})}|W(H_{0,1},T_1)|^{-1}\int_{T_1(F)/Z_{G_1}(F)}D^{G_1}(t_1)^{1/2}\Delta(t_1)^{s-1/2}c_{\pi_2}(t_1)\omega_1(t_1)^{-1}\,\mathrm{d}_vt_1.$$

This proves (5.6). □

Now we are ready to prove Theorem 1.2. The argument is similar to the unitary similitude group case. For $i=1,2$, fix $\varepsilon_i\in F^\times$ with $\eta_{E/F}(\varepsilon_i)=(-1)^{i-1}$ as before. Let $\Pi_\phi=\Pi_\phi(G_{1,\varepsilon_1})\cup\Pi_\phi(G_{1,\varepsilon_2})$ be a local tempered Vogan L-packet of $\mathrm{U}_6(F)$. Our goal is to show that

$$\sum_{\pi_1\in\Pi_\phi}m(\pi_1)=2. \tag{5.8}$$

By the multiplicity formula in Proposition 5.5, we have

$$\sum_{\pi_1\in\Pi_\phi}m(\pi_1)=\sum_{i=1}^{2}\sum_{\pi_{1,\varepsilon_i}\in\Pi_\phi(G_{1,\varepsilon_i})}c_{\pi_{1,\varepsilon_i}}(1)+\sum_{i=1}^{2}\sum_{\pi_{1,\varepsilon_i}\in\Pi_\phi(G_{1,\varepsilon_i})}\sum_{T_i\in\mathcal{T}_{ell}(H_{0,1,\varepsilon_i})}$$
$$|W(H_{0,1,\varepsilon_i},T_i)|^{-1}\int_{T_i(F)/Z_{G_{1,\varepsilon_i}}(F)}D^{H_{1,\varepsilon_i}}(t_i)c_{\pi_{1,\varepsilon_i}}(t_i)\omega_{1,\varepsilon_i}^{-1}(t_i)\,\mathrm{d}t_i.$$

By the same argument as in the proof of Theorem 4.13, together with Theorem 2.1 (2) and (3), we can show that

$$\sum_{i=1}^{2}\sum_{\pi_{1,\varepsilon_i}\in\Pi_\phi(G_{1,\varepsilon_i})}c_{\pi_{1,\varepsilon_i}}(1)=1,$$

$$\sum_{i=1}^{2}\sum_{\pi_{1,\varepsilon_i}\in\Pi_\phi(G_{1,\varepsilon_i})}\sum_{T_i\in\mathcal{T}_{ell}(H_{0,1,\varepsilon_i})}|W(H_{0,1,\varepsilon_i},T_i)|^{-1}\int_{T_i(F)/Z_{G_{1,\varepsilon_i}}(F)}D^{H_{1,\varepsilon_i}}(t_i)c_{\pi_{1,\varepsilon_i}}(t_i)\omega_{1,\varepsilon_i}^{-1}(t_i)\,\mathrm{d}t_i=0.$$

This proves (5.8) and finishes the proof of Theorem 1.2.

To summarize, we have reduced the proofs of the main theorems (i.e. Theorems 1.1 and 1.2) to the proof of the trace formula in Theorem 4.3.

6. The proof of the spectral side of the trace formula

In this section, we will prove the trace formula. Since the geometric side of the trace formula has already been proved in Section 4.3, we only need to consider the spectral side. As in the previous sections, we will use (G, H) to denote $(G_\varepsilon, H_\varepsilon)$ and use $\omega \otimes \xi$ to denote $\omega_\varepsilon \otimes \xi_\varepsilon$.

6.1. **Explicit intertwining operator $\mathcal{L}_\pi$.** By Lemma 3.3, for all $f \in \mathcal{C}(Z_G(F)\backslash G(F), \eta^{-1})$, the integral

$$\int_{Z_H(F)\backslash H(F)} f(h)\omega \otimes \xi(h)\,\mathrm{d}h$$

is absolutely convergent and defines a continuous linear form on the space $\mathcal{C}(Z_G(F)\backslash G(F), \eta^{-1})$.

Proposition 6.1. *(1) The linear form*

$$f \in \mathcal{C}(Z_G(F)\backslash G(F), \eta^{-1}) \to \int_{Z_H(F)\backslash H(F)} f(h)\omega \otimes \xi(h)\,\mathrm{d}h$$

can be extended continuously to $\mathcal{C}^w(Z_G(F)\backslash G(F), \eta^{-1})$. We will use

$$\mathcal{P}_{H,\omega\otimes\xi} : \ f \in \mathcal{C}^w(Z_G(F)\backslash G(F), \eta^{-1}) \to \int^*_{Z_H(F)\backslash H(F)} f(h)\omega \otimes \xi(h)\,\mathrm{d}h$$

to denote this continuous linear form.

(2) For all $f \in \mathcal{C}^w(Z_G(F)\backslash G(F), \eta^{-1})$, and $h_0, h_1 \in H(F)$, we have

$$\mathcal{P}_{H,\omega\otimes\xi}(L(h_0)R(h_1)f) = \omega \otimes \xi(h_0 h_1^{-1})\mathcal{P}_{H,\omega\otimes\xi}(f)$$

where R (resp. L) is the right (resp. left) translation.

Proof. The proof is very similar to the GR case (Proposition 5.1 and Lemma 5.2 of [18]), and we will skip it here. □

Lemma 6.2. *(1) Let π be an irreducible tempered representation of $G(F)$ with central character η and let $l \in \mathrm{Hom}_H(\pi, \omega \otimes \xi)$ be a continuous $(H, \omega \otimes \xi)$-equivariant linear form. Then there exist $d > 0$ and a continuous semi-norm ν_d on π such that*

$$|l(\pi(x)e)| \leq \nu_d(e)\Xi^{H\backslash G}(x)\sigma_{H\backslash G}(x)^d$$

for all $e \in \pi$ and $x \in H(F)\backslash G(F)$.

(2) For all $d > 0$, there exist $d' > 0$ and a continuous semi-norm $\nu_{d,d'}$ on $\mathcal{C}_d^w(Z_G(F)\backslash G(F), \eta^{-1})$ such that

$$|\mathcal{P}_{H,\xi}(R(x)L(y)\varphi)| \leq \nu_{d,d'}(\varphi)\Xi^{H\backslash G}(x)\Xi^{H\backslash G}(y)\sigma_{H\backslash G}(x)^{d'}\sigma_{H\backslash G}(y)^{d'}$$

for all $\varphi \in \mathcal{C}_d^w(Z_G(F)\backslash G(F), \eta^{-1})$ and $x, y \in H(F)\backslash G(F)$.

Proof. The proof is the same as the GR case (Proposition 6.2.3 of [20]), we will skip it here. □

Let π be a tempered representation of $G(F)$ with central character η and $\mathrm{End}(\pi)$ be the space of continuous endomorphisms of the space of π. It is a continuous representation of $G(F)\times G(F)$ via the left and right translation. As in Section 2.2 of [2], let $\mathrm{End}(\pi)^{\infty}$ be the subspace of smooth vectors in $\mathrm{End}(\pi)$. For all $T\in \mathrm{End}(\pi)^{\infty}$, define

$$\mathcal{L}_{\pi}(T)=\mathcal{P}_{H,\xi}(\mathrm{tr}(\pi(g^{-1})T))=\int^{*}_{Z_H(F)\backslash H(F)}\mathrm{tr}(\pi(h^{-1})T)\omega\otimes\xi(h)\,\mathrm{d}h.$$

The map $\mathcal{L}_{\pi}:\mathrm{End}(\pi)^{\infty}\to\mathbb{C}$ is a linear form in $\mathrm{End}(\pi)^{-\infty}$. Here $\mathrm{End}(\pi)^{-\infty}$ is the dual of $\mathrm{End}(\pi)^{\infty}$. By Proposition 6.1, for any $h,h'\in H(F)$, we have

$$\mathcal{L}_{\pi}(\pi(h)T\pi(h'))=\omega\otimes\xi(hh')\mathcal{L}_{\pi}(T). \tag{6.1}$$

For $e,e'\in\pi$, define $T_{e,e'}\in \mathrm{End}(\pi)^{\infty}$ to be the map $e_0\in\pi\mapsto(e_0,e')e$. Set $\mathcal{L}_{\pi}(e,e')=\mathcal{L}_{\pi}(T_{e,e'})$. Then

$$\mathcal{L}_{\pi}(e,e')=\int^{*}_{Z_H(F)\backslash H(F)}(e,\pi(h)e')\omega\otimes\xi(h)\,\mathrm{d}h.$$

If we fix e', by (6.1), the map $e\in\pi\to\mathcal{L}_{\pi}(e,e')$ belongs to $Hom_H(\pi,\omega\otimes\xi)$. Since $Span\{T_{e,e'}\colon e,e'\in\pi\}=\mathrm{End}(\pi)^{\infty}$, we have

$$\mathcal{L}_{\pi}\neq 0\Rightarrow m(\pi)\neq 0. \tag{6.2}$$

Later in this subsection, we will show that the other direction also holds. Before that, we discuss some basic properties of $\mathcal{L}_{\pi}$.

Lemma 6.3. *With the notation above, the following holds.*

(1) The map $\pi\in\Pi_{temp}(G,\eta)\to\mathcal{L}_{\pi}\in\mathrm{End}(\pi)^{-\infty}$ is smooth in the sense of Lemma 7.2.2(i) of [2].

(2) For $f\in\mathcal{C}(Z_G(F)\backslash G(F),\eta^{-1})$, we have

$$\int_{Z_H(F)\backslash H(F)}f(h)\omega\otimes\xi(h)\,\mathrm{d}h=\int_{\Pi_{temp}(G,\eta)}\mathcal{L}_{\pi}(\pi(f))\mu(\pi)\,\mathrm{d}\pi$$

with both integrals being absolutely convergent.

(3) For $f\in\mathcal{C}_{ind}(Z_G(F)\backslash G(F),\eta^{-1})$ and $f'\in\mathcal{C}(Z_G(F)\backslash G(F),\eta)$, we have

$$\int_{\Pi_{temp}(G,\eta)}\mathcal{L}_{\pi}(\pi(f))\overline{\mathcal{L}_{\pi}(\pi(\bar{f}'))}\mu(\pi)\,\mathrm{d}\pi$$

$$=\int_{Z_H(F)\backslash H(F)}\int_{Z_H(F)\backslash H(F)}\int_{Z_G(F)\backslash G(F)}f(hgh')f'(g)\,\mathrm{d}g\omega\otimes\xi(h')\,\mathrm{d}h'\omega\otimes\xi(h)\,\mathrm{d}h$$

where the left hand side is absolutely convergent and the right hand side is convergent in that order but is not necessarily absolutely convergent.

Proof. The proof is similar to the GGP case (Lemma 7.2.2 of [2]) and the GR case (Lemma 6.2.2 of [20]), so we will skip it here. The only thing we want to point out is that the proof of (3) in both previous cases uses the Gelfand pair condition (i.e. $m(\pi) \leq 1$ for all irreducible tempered representations π). But we don't have this condition for the current model, all we have is that $m(\pi) \leq 1$ for all $\pi \in \Pi_{temp}(G,\eta) \smallsetminus \Pi_2(G,\eta)$ (see Corollary 4.15). This is why we require $f \in \mathcal{C}_{ind}(Z_G(F)\backslash G(F), \eta^{-1})$ in the statement of (3) (note that in the previous two cases, the equality in (3) holds for all $f \in \mathcal{C}(Z_G(F)\backslash G(F), \eta^{-1})$). □

We then study the behavior of $\mathcal{L}_\pi$ under the parabolic induction. Let $\bar{Q} = M_Q\bar{U}_Q$ be a good parabolic subgroup, and $\tau \in \Pi_2(M_Q)$ be a discrete series whose central character equals η on $Z_G(F)$. Set $\pi = I_{\bar{Q}}^G(\tau)$. Then π is a tempered representation of G with central character η. Let $H_{\bar{Q}} = H \cap \bar{Q}$. For $T \in \mathrm{End}(\tau)^\infty$, define

$$\mathcal{L}_\tau(T) = \int_{Z_H(F)\backslash H_{\bar{Q}}(F)} \mathrm{tr}(\tau(h_{\bar{Q}}^{-1})T)\delta_{H_{\bar{Q}}}(h_{\bar{Q}})^{1/2}\omega\otimes\xi(h_{\bar{Q}})\,\mathrm{d}h_{\bar{Q}}.$$

The integral above is absolutely convergent by Proposition 3.5(2) together with the assumption that τ is a discrete series.

Proposition 6.4. *With the notation above, we have*

$$\mathcal{L}_\pi \neq 0 \iff \mathcal{L}_\tau \neq 0.$$

Proof. The proof is very similar to the GR model case (Proposition 5.6 of [18]), we will skip it here. □

The following proposition tells us the relation between $\mathcal{L}_\pi$ and $m(\pi)$.

Proposition 6.5. *Let π be an irreducible tempered representation of $G(F)$ with central character η. Then $\mathcal{L}_\pi \neq 0 \iff m(\pi) \neq 0$.*

Proof. By (6.2), we only need to show that $m(\pi) \neq 0 \Rightarrow \mathcal{L}_\pi \neq 0$. From now on, we assume that $m(\pi) \neq 0$. Fix $0 \neq l \in Hom_{H(F)}(\pi, \omega\otimes\xi)$. For $T \in C_c^\infty(\Pi_{temp}(G,\eta))$, by the same argument as in the GR model case (Section 5.5 of [18]), we have

$$l(T_\pi e) = \int_{H(F)\backslash G(F)} l(\pi(x)e) \int_{\Pi_{temp}(G,\eta)} \mathcal{L}_\Pi(T_\Pi \Pi(x^{-1}))\mu(\Pi)\,\mathrm{d}\Pi\,\mathrm{d}x \tag{6.3}$$

for all $e \in \pi$. Choose a good parabolic subgroup $\bar{Q} = M_Q\bar{U}_Q$ of G and $\tau \in \Pi_2(M_Q)$ such that π is a direct summand of $\pi' = I_{\bar{Q}}^G(\tau)$. Let

$$\mathcal{O} = \{I_{\bar{Q}}^G(\tau_\lambda) \colon \lambda \in i\mathfrak{a}_{M_Q,0}^*\} \subset \Pi_{temp}(G,\eta) \tag{6.4}$$

be the connected component containing π'. Choose $e_0 \in \pi$ such that $l(e_0) \neq 0$, and let $T_0 \in \mathrm{End}(\pi)^\infty$ with $T_0(e_0) = e_0$. We can easily find an element $T^0 \in C_c^\infty(\Pi_{temp}(G,\eta))$ such that

$$T_\pi^0 = T_0,\ Supp(T^0) \subset \mathcal{O}.$$

By applying (6.3) to the case $e = e_0$ and $T = T^0$, we know there exists $\lambda \in i\mathfrak{a}_{M_Q,0}^*$ such that $\mathcal{L}_{\pi'_\lambda} \neq 0$ where $\pi'_\lambda = I_{\bar{Q}}^G(\tau_\lambda)$. By Proposition 6.4, this implies that $\mathcal{L}_{\tau_\lambda} \neq 0$. We need a lemma:

Lemma 6.6. *For all $\lambda \in i\mathfrak{a}_{M_Q,0}^*$, we have*

$$\mathcal{L}_\tau \neq 0 \iff \mathcal{L}_{\tau_\lambda} \neq 0.$$

Proof. When the reduced model $(G_{\bar{Q}}, H_{\bar{Q}})$ is of Type I (defined in Section 4.5), it is easy to see from the definition that the nonvanishing property of $\mathcal{L}_\tau$ is invariant under the unramified twist.

When the reduced model is of Type II, it is not clear from the definition that the unramified twist will preserve the nonvanishing property. Instead, we claim that $\mathcal{L}_\tau$ is always nonzero in this case. In fact, since τ is a discrete series, by the same argument as in the GR model case (i.e. Remark 5.12 of [18]), we have $m(\tau) \neq 0 \Rightarrow \mathcal{L}_\tau \neq 0$. Hence it is enough to show that the multiplicity is always nonzero for Type II reduced models. But this just follows from Theorem 4.13. □

Back to the proof of the proposition. The lemma above implies that $\mathcal{L}_\tau \neq 0$. Together with Proposition 6.4, we have $\mathcal{L}_{\pi'} \neq 0$. Since π is a direct summation of π', we can write π' as $\pi \oplus \pi_0$. If $\mathcal{L}_{\pi_0} \neq 0$, then $m(\pi_0) \neq 0$ which implies that $m(\pi') = m(\pi) + m(\pi_0) \geq 2$. Meanwhile, by Corollary 4.15, we know that $m(\pi') \leq 1$. So we get a contradiction. Hence we have $\mathcal{L}_{\pi_0} = 0$. But since $\mathcal{L}_{\pi'} \neq 0$, we have $\mathcal{L}_\pi \neq 0$. This finishes the proof of the proposition. □

Now we are ready to prove Theorem 4.16 by assuming Proposition 4.14 holds. We first recall the statement of the theorem. Let $\bar{Q} = M_Q\bar{U}_Q$ be a good parabolic subgroup of G, and let τ be a smooth finite length tempered representation of $M_Q(F)$ whose central character equals η on $Z_G(F)$. Set $\pi = I_{\bar{Q}}^G(\tau)$. Our goal is to show that $m(\pi) = m(\tau)$. Without loss of generality, we can assume that τ is irreducible. By Theorem 4.13 and Proposition 4.14, it is enough to show that

$$m(\tau) \neq 0 \Rightarrow m(\pi) \neq 0. \tag{6.5}$$

There exists a good parabolic subgroup $\bar{Q}' = M'_Q\bar{U}'_Q$ of G, and an irreducible discrete series σ of $M'_Q(F)$ such that τ is a direct summand of $I_{\bar{Q}'\cap M'_Q}^{M_Q}(\sigma)$. Let $I_{\bar{Q}'\cap M'_Q}^{M_Q}(\sigma) = \tau \oplus \tau_0$ and $I_{\bar{Q}'}^G(\sigma) = \pi \oplus \pi_0$. Then $\pi_0 = I_{\bar{Q}}^G(\tau_0)$. By Theorem 4.13 and an analogue of Proposition 4.14 for the reduced models (see Appendix A.1), we have $m(\tau) + m(\tau_0) \leq m(\sigma) \leq 1$. Since $m(\tau) \neq 0$, we have $m(\tau_0) = 0$ and $m(\sigma) \neq 0$. Applying Proposition 4.14 again, we have $m(\pi_0) = 0$.

Since $m(\sigma) \neq 0$, we have $\mathcal{L}_\sigma \neq 0$. By Proposition 6.4, we know that $\mathcal{L}_{\pi\oplus\pi_0} \neq 0$. But since $m(\pi_0) = 0$, by Proposition 6.5, we know that $\mathcal{L}_{\pi_0} = 0$. This implies that $\mathcal{L}_\pi \neq 0$ and hence $m(\pi) \neq 0$. This proves (6.5) and finishes the proof of Theorem 4.16.

To end this subsection, we prove a proposition that will be used later in the proof of the spectral side of the trace formula.

Proposition 6.7. *Let $\mathcal{K} \subset \Pi_{temp}(G,\eta)$ be a compact subset. Then there exists an element $T \in \mathcal{C}(\Pi_{temp}(G,\eta))$ such that $\mathcal{L}_\pi(T_\pi) = m(\pi)$ for all $\pi \in \mathcal{K}$.*

Proof. It is enough to show that for all $\pi' \in \Pi_{temp}(G,\eta)$, there exists $T \in \mathcal{C}(\Pi_{temp}(G,\eta))$ such that $\mathcal{L}_\pi(T_\pi) = m(\pi)$ for all π in some neighborhood of π' in $\Pi_{temp}(G,\eta)$. If π' is a discrete series, by Theorem 6.5, we can find $T' \in \mathrm{End}(\pi')^\infty$ such that $\mathcal{L}_{\pi'}(T') = m(\pi')$. Then we just need to take any $T \in \mathcal{C}(\Pi_{temp}(G,\eta))$ with $T_{\pi'} = T'$.

If π' is not a discrete series, we can find a good parabolic subgroup $\bar{Q} = M_Q\bar{U}_Q$ of G and $\tau \in \Pi_2(M_Q)$ such that $\pi' = I_{\bar{Q}}^G(\tau)$. Let $\mathcal{O} = \{I_{\bar{Q}}^G(\tau_\lambda) \colon \lambda \in i\mathfrak{a}_{M_Q,0}^*\}$ be the connected component containing π'. We first show that

(1): The multiplicity is constant on $\mathcal{O}$ (i.e. $m(\pi) = m(\pi')$ for all $\pi \in \mathcal{O}$).

By Theorem 4.16, it is enough to show that the multiplicity $m(\tau)$ of the reduced model $(G_{\bar{Q}}, H_{\bar{Q}})$ is invariant under unramified twist. When the reduced model is of Type I, this just follows from the definition. If the reduced model is of Type II, by Theorem 4.13, the multiplicity is always equal to 1 and hence invariant under the unramified twist. This proves (1).

Now if $m(\pi') = 0$, then $m(\pi) = 0$ for all $\pi \in \mathcal{O}$ and we can just take $T = 0$. If $m(\pi') \neq 0$, by the discussion above together with Corollary 4.15, we know that $m(\pi) = 1$ for all $\pi \in \mathcal{O}$. By Theorem 6.5, we can find $T' \in \mathrm{End}(\pi')^\infty$ such that $\mathcal{L}_{\pi'}(T') \neq 0$. Let $T^0 \in \mathcal{C}(\Pi_{temp}(G,\eta))$ be an element with $T^0_{\pi'} = T'$. By Lemma 6.3(1), the function $\pi \to \mathcal{L}_\pi(T^0_\pi)$ is a smooth function. The value at π' is just $\mathcal{L}_{\pi'}(T') \neq 0$. Hence we can find a smooth compactly supported function φ on $\Pi_{temp}(G,\eta)$ such that $\varphi(\pi)\mathcal{L}_\pi(T^0_\pi) = 1$ for all π belonging to a small neighborhood of π'. Then we just need to take $T = \varphi T^0$ and this finishes the proof of the Proposition. □

6.2. **The proof of the spectral side of the trace formula.** In this subsection, we will prove the spectral side of the trace formula. Since we only know $m(\pi) \leq 1$ for $\pi \in \Pi_{temp}(G) \smallsetminus \Pi_2(G)$, we need to divide the proof into two steps.

Step 1: We first prove the spectral side of the trace formula for $f \in {}^\circ\mathcal{C}(Z_G(F)\backslash G(F), \eta^{-1})$. For such f, the spectral side only contains discrete series. By a similar argument as in the Galois model case (Section 3 of [3]), we can prove the trace formula without using the multiplicity one assumption.

Step 2: We then prove the trace formula for $f \in \mathcal{C}_{ind,scusp}(Z_G(F)\backslash G(F), \eta^{-1})$. For this kind of test functions, the spectral side will only contain tempered representations which are not discrete series. But for these representations, the multiplicity one assumption holds by Corollary 4.15. Hence we can prove the trace formula by the same argument as in the GGP case and the GR case.

We start with Step 1. For $\pi \in \Pi_2(G,\eta)$, let

$$\mathcal{B}_\pi : \pi \times \pi^\vee \to \mathbb{C}$$

be the bilinear form defined by

$$\mathcal{B}_\pi(v,v^\vee) := \int_{Z_H(F)\backslash H(F)} \langle v, \pi(h)v^\vee\rangle(\omega\otimes\xi)(h)\,\mathrm{d}h,\ (v,v^\vee)\in \pi\times\pi^\vee.$$

Note that the integral above is absolutely convergent by Lemma 3.3. It is easy to see from the definition that $\mathcal{B}_\pi$ descents to a bilinear pairing

$$\mathcal{B}_\pi : \pi_{\omega\otimes\xi} \times \pi^\vee_{(\omega\otimes\xi)^{-1}} \to \mathbb{C}$$

where $\pi_{\omega\otimes\xi}$ (resp. $\pi^\vee_{(\omega\otimes\xi)^{-1}}$) is the $(H,\omega\otimes\xi)$-coinvariant spaces (resp. $(H,(\omega\otimes\xi)^{-1})$-coinvariant) of π (resp. $\pi^\vee$).

Proposition 6.8. *$\mathcal{B}_\pi$ induces a perfect pairing between $\pi_{\omega\otimes\xi}$ and $\pi^\vee_{(\omega\otimes\xi)^{-1}}$.*

Proof. This proposition follows from exactly the same argument as Proposition 3.2.1 of [3] after we establish the next lemma.

Lemma 6.9. *For all $l \in Hom_H(\pi,\omega\otimes\xi)$, $v\in\pi$ and $f\in\mathcal{C}(Z_G(F)\backslash G(F),\eta^{-1})$, the integrals*

$$\int_{H(F)\backslash G(F)} |l(\pi(x)v)|^2\,\mathrm{d}x,\quad \int_{Z_G(F)\backslash G(F)} f(g)l(\pi(g)v)\,\mathrm{d}g$$

are absolutely convergent. Moreover, we have

$$l(\pi(f)v) = \int_{Z_G(F)\backslash G(F)} f(g)l(\pi(g)v)\,\mathrm{d}g. \tag{6.6}$$

Proof. For any $d>0$ and $e\in\pi$, by a similar argument as in the GGP case (Lemma 7.3.1 of [2]) or the GR case (Lemma 6.2.3 of [20]), we have

$$|l(\pi(x)e)| \ll \Xi^{H\backslash G}(x)\sigma_{H\backslash G}(x)^{-d} \tag{6.7}$$

for all $x\in H(F)\backslash G(F)$. Note that both loc. cit. considered all tempered representations, and hence the right hand side in both loc. cit. is $\Xi^{H\backslash G}(x)\sigma_{H\backslash G}(x)^d$. Here since we only consider discrete series, we can have better bound on the right hand side (i.e. $\Xi^{H\backslash G}(x)\sigma_{H\backslash G}(x)^{-d}$). Combining (6.7) with Proposition 3.4, we know that both integrals in the lemma are absolutely convergent. Finally, (6.6) follows from the standard argument as in the GGP case (Section 7.5 of [2]) and the GR case (Section 6.5 of [20]). □

□

Now we are ready to prove the spectral side of the trace formula for $f\in {}^\circ\mathcal{C}(Z_G(F)\backslash G(F),\eta^{-1})$. We fix such a test function f. Without loss of generality, we may assume that there exist $\pi\in\Pi_2(G,\chi)$, $v\in\pi$ and $v^\vee\in\pi^\vee$ such that $f(g)=\langle v,\pi(g)v^\vee\rangle$. Then the spectral side becomes

$$I_{spec}(f) = \mathrm{tr}(\pi(f))m(\bar{\pi}) = \mathrm{tr}(\pi(f))m(\pi).$$

For $x,y\in G(F)$, define

$$K_f(x,y)=\int_{Z_H(F)\backslash H(F)} f(x^{-1}hy)\omega\otimes\xi(h)\,\mathrm{d}h.$$

Then we have

$$K_f(x,y)=\mathcal{B}_\pi(\pi(x)v,\pi^\vee(y)v') \text{ and } K_f(x,x)=I(f,x).$$

Let $N=m(\pi)$ and let $v_1,\cdots,v_N$ be vectors in π whose images in $\pi_{\omega\otimes\xi}$ form a basis. We then let $v_1^\vee,\cdots,v_N^\vee$ be vectors in $\pi^\vee$ whose images in $\pi^\vee_{(\omega\otimes\xi)^{-1}}$ form the dual basis under the pairing $\mathcal{B}_\pi$. As in the Galois model case (Section 3 of [3]), we have

$$\begin{aligned}
I(f) &= \int_{H(F)\backslash G(F)} I(f,x)\,\mathrm{d}x=\int_{H(F)\backslash G(F)} \mathcal{B}_\pi(\pi(x)v,\pi^\vee(x)v')\,\mathrm{d}x\\
&= \sum_{i=1}^N\int_{H(F)\backslash G(F)} \mathcal{B}_\pi(\pi(x)v,v_i^\vee)\mathcal{B}_\pi(v_i,\pi^\vee(x)v')\,\mathrm{d}x\\
&= \sum_{i=1}^N\int_{Z_G(F)\backslash G(F)} \langle\pi(g)v,v_i^\vee\rangle\mathcal{B}_\pi(v_i,\pi^\vee(g)v^\vee)\,\mathrm{d}g\\
&= \sum_{i=1}^N\frac{\langle v,v^\vee\rangle}{d(\pi)}\mathcal{B}_\pi(v_i,v_i^\vee)=\mathrm{tr}(\pi(f))\cdot N=\mathrm{tr}(\pi(f))m(\pi)
\end{aligned}$$

where $d(\pi)$ is the formal degree of π. This proves the spectral side of the trace formula for all $f\in {}^\circ\mathcal{C}(Z_G(F)\backslash G(F),\eta^{-1})$.

Now it remains to consider the case when $f\in\mathcal{C}_{ind,scusp}(Z_G(F)\backslash G(F),\eta^{-1})$. We fix such a function f. Then the spectral side does not contain discrete series. By Corollary 4.15, we have the Gelfand pair condition for all the representations appear in the spectral side. As a result, we can prove the trace formula by a similar argument as in the GGP case and the GR case. To be specific, for $f'\in\mathcal{C}(Z_G(F)\backslash G(F),\eta)$, define

$$K_{f,f'}(g_1,g_2)=\int_{Z_G(F)\backslash G(F)} f(g_1^{-1}gg_2)f'(g)\,\mathrm{d}g,\ g_1,g_2\in G(F),$$

$$K_{f,f'}^H(x,y)=\int_{Z_H(F)\backslash H(F)}\int_{Z_H(F)\backslash H(F)} K_{f,f'}(h_1^{-1}x,h_2y)\omega\otimes\xi(h_1h_2)\,\mathrm{d}h_1\,\mathrm{d}h_2,\ x,y\in G(F),$$

$$J_{aux}(f,f')=\int_{H(F)\backslash G(F)} K_{f,f'}^H(x,x)\,\mathrm{d}x.$$

By the same argument as in Proposition 7.5 of [17], we know that the integral defining $K_{f,f'}$ is absolutely convergent, the two integrals defining $K_{f,f'}^H$ are absolutely convergent (although the double integral is not necessarily absolutely convergent), and the integral defining $J_{aux}(f,f')$ is absolutely convergent.

Proposition 6.10. *We have*

$$K_{f,f'}^{H}(x,y)=\int_{\Pi_{temp}(G,\eta)}\mathcal{L}_{\pi}(\pi(x)\pi(f)\pi(y^{-1}))\overline{\mathcal{L}_{\pi}(\pi(\overline{f'}))}\mu(\pi)\,\mathrm{d}\pi, \tag{6.8}$$

$$J_{aux}(f,f')=\int_{\mathcal{X}(G,\eta)}D(\pi)\theta_f(\pi)\overline{\mathcal{L}_{\pi}(\pi(\overline{f'}))}\,\mathrm{d}\pi. \tag{6.9}$$

Proof. The proof is very similar to the GGP case (Proposition 9.2.1 and 9.2.2 of [2]) and the GR case (Propositions 8.2.2 and 8.2.3 of [20]), we will skip it here. The only thing we want to point out is that the proof of (6.8) uses Lemma 6.3(3) whose proof uses the Gelfand pair condition. This is why we need to require $f\in\mathcal{C}_{ind,scusp}(Z_G(F)\backslash G(F),\eta^{-1})$. □

Now we are ready to prove the trace formula. By Lemma 6.3(2), together with the fact that $f\in\mathcal{C}_{ind,scusp}(Z_G(F)\backslash G(F),\eta^{-1})$, we have

$$I(f,x)=\int_{\Pi_{temp}(G,\eta)\smallsetminus\Pi_2(G,\eta)}\mathcal{L}_{\pi}(\pi(x)\pi(f)\pi(x)^{-1})\mu(\pi)\,\mathrm{d}\pi. \tag{6.10}$$

By Proposition 6.7, there exists a function $f'\in\mathcal{C}(Z_G(F)\backslash G(F),\eta)$ such that

$$\mathcal{L}_{\pi}(\pi(\overline{f'}))=m(\pi)$$

for all $\pi\in\Pi_{temp}(G,\eta)$ with $\pi(f)\neq 0$ (note that $\pi(f)\neq 0$ will imply $\pi\in\Pi_{temp}(G,\eta)\smallsetminus\Pi_2(G,\eta)$). We fix such a function f'. By Theorem 6.5 and Corollary 4.15, for all $\pi\in\Pi_{temp}(G,\eta)\smallsetminus\Pi_2(G,\eta))$, $\mathcal{L}_{\pi}\neq 0$ if and only if $m(\pi)=1$. Then (6.10) becomes

$$I(f,x)=\int_{\pi\in\Pi_{temp}(G,\eta)\smallsetminus\Pi_2(G,\eta))}\mathcal{L}_{\pi}(\pi(x)\pi(f)\pi(x)^{-1})\overline{\mathcal{L}_{\pi}(\pi(\overline{f'}))}\mu(\pi)\,\mathrm{d}\pi.$$

Combining with Proposition 6.10, we have $I(f,x)=K_{f,f'}^{H}(x,x)$. Therefore $I(f)=J_{aux}(f,f')$. Applying Proposition 6.10 again, together with the fact that $\overline{\mathcal{L}_{\pi}(\pi(\overline{f'}))}=m(\pi)=m(\bar{\pi})$, we have

$$I(f)=J_{aux}(f,f')=\int_{\mathcal{X}(G,\eta)}D(\pi)\theta_f(\pi)m(\bar{\pi})\,\mathrm{d}\pi=I_{spec}(f).$$

This finishes the proof of the spectral side of the trace formula.

Appendix A. The proof of Proposition 4.14

A.1. **Some reduction.** We first recall the notation and the statement of the proposition. Let $\bar{Q}=M_Q\bar{U}_Q$ be a good parabolic subgroup of G, and τ be a smooth finite length tempered representation of $M_Q(F)$ whose central character equals η on $Z_G(F)$. Set $\pi=I_{\bar{Q}}^{G}(\tau)$. Our goal is to show that

$$m(\pi)\leq m(\tau). \tag{A.1}$$

Here $m(\pi)$ is the multiplicity for the model (G,H) and $m(\tau)$ is the multiplicity for the reduced model $(G_{\bar{Q}},H_{\bar{Q}})$. We will apply Mackey theory to prove (A.1). To simplify our notation, we assume that the characters η and

ω are trivial. The argument for the general case (i.e. when the characters are nontrivial) will be exactly the same as this case.

First we consider the analogue of (A.1) for reduced models. Let $M_Q\bar{U}_Q = \bar{Q} \subset \bar{Q}' = M_Q'\bar{U}_Q'$ be two good parabolic subgroups, and let $(G_{\bar{Q}'}, H_{\bar{Q}'})$, $(G_{\bar{Q}}, H_{\bar{Q}})$ be the associated reduced models. Given a smooth finite length tempered representation τ of $M_{\bar{Q}}(F)$, set $\tau' = I_{\bar{Q}\cap M_Q'}^{M_Q'}(\tau)$ which is a smooth finite length tempered representation of $M_{\bar{Q}}'(F)$. Let $m(\tau)$ and $m(\tau')$ be the multiplicity of the reduced models $(G_{\bar{Q}}, H_{\bar{Q}})$ and $(G_{\bar{Q}'}, H_{\bar{Q}'})$ respectively. Then the analogue of (A.1) for reduced models is just

$$m(\tau') \leq m(\tau). \tag{A.2}$$

The proof of (A.2) follows from the same, but easier arguments as the proof of (A.1). Hence by induction, we will assume that (A.2) holds for all good parabolic subgroups $\bar{Q} \subset \bar{Q}'$.

Then we reduce the proof of (A.1) to the case when $\bar{Q}$ is a maximal parabolic subgroup. In general, if $\bar{Q}$ is not a maximal parabolic subgroup, we can find a good maximal parabolic subgroup $\bar{Q}' = M_Q'\bar{U}_Q'$ with $\bar{Q} \subset \bar{Q}'$. Let $\tau' = I_{\bar{Q}\cap M_Q'}^{M_Q'}(\tau)$ be a smooth finite length tempered representation of $M_Q'(F)$ and let $m(\tau')$ be the multiplicity for the reduced model $(G_{\bar{Q}'}, H_{\bar{Q}'})$. Then we have $\pi = I_{\bar{Q}'}^{G}(\tau')$. Once we have proved (A.1) for all maximal parabolic subgroups, we have $m(\pi) \leq m(\tau')$. By (A.2), we also have $m(\tau') \leq m(\tau)$. This proves (A.1) for the parabolic subgroup $\bar{Q}$. So from now on, we can assume that $\bar{Q}$ is a maximal good parabolic subgroup. We can also assume that τ is irreducible.

Since G/H is a spherical variety, the double coset $\bar{Q}(F)\backslash G(F)/H(F)$ only contains finitely many elements (Section 0.2 of [4] and Lemma 3.4.1 of [12]), and we denote it by $\{\bar{Q}\gamma_i H\colon 1 \leq i \leq k\}$. By the geometric lemma of Bernstein-Zelevinsky in [6], we may reorder γ_i such that

$$Y_i = \cup_{j=1}^{i} \bar{Q}(F)\gamma_j H(F)$$

is an open subset of $G(F)$ for all $1 \leq i \leq k$. Since $\bar{Q}$ is a good parabolic subgroup, we may assume that $\gamma_1 = 1$.

With the filtration above, for $1 \leq i \leq k$, define

$$V_i = \{f \in I_{\bar{Q}}^{G}(\tau)\colon supp(f) \subset Y_i\}.$$

Then we have $V_1 \subset V_2 \subset \cdots \subset V_k = \pi$ and V_i is $H(F)$-invariant for all i. In particular, this implies that

$$m(\pi) = \dim(\mathrm{Hom}_{H(F)}(\pi, \xi)) \leq \sum_{i=1}^{k} \dim(\mathrm{Hom}_{H(F)}(V_i/V_{i-1}, \xi)). \tag{A.3}$$

Here $V_0 = \{0\}$. Moreover, for any $1 \leq i \leq k$, it is easy to see that the map

$$f \in V_i \mapsto \phi_f(h) := f(y_i h)$$

is an isomorphism between V_i/V_{i-1} and $ind_{H_i}^H(\delta_{\bar{Q}}^{1/2}\tau^{\gamma_i}|_{H_i})$ ($ind_{H_i}^H$ is the compact induction). Here $H_i = H(F)\cap\gamma_i^{-1}\bar{Q}(F)\gamma_i = \gamma_i^{-1}Q_i(F)\gamma_i$ with $Q_i(F) = \bar{Q}(F)\cap\gamma_i H(F)\gamma_i^{-1}$. By reciprocity law, we have

$$\mathrm{Hom}_{H(F)}(V_i/V_{i-1},\xi)\simeq \mathrm{Hom}_{Q_i(F)}(\tau,(\delta_{Q_i}\delta_{\bar{Q}}^{-1/2})\otimes{}^{\gamma_i}\xi) \tag{A.4}$$

where the character ${}^{\gamma_i}\xi$ is defined by ${}^{\gamma_i}\xi(q) := \xi(\gamma_i^{-1}q\gamma_i)$ for $q\in Q_i(F)$. Here we view τ as a representation of $\bar{Q}(F)$ by making it trivial on $\bar{U}_Q(F)$.

Proposition A.1. *With the notation above, we have*

$$m(\tau) = \dim(\mathrm{Hom}_{H(F)}(V_1,\xi)).$$

Proof. Since $\gamma_1 = 1$, we have $Q_1 = H\cap\bar{Q} = H_{\bar{Q}}$. By (A.4), we have

$$\mathrm{Hom}_{H(F)}(V_1,\xi)\simeq \mathrm{Hom}_{H_{\bar{Q}}(F)}(\tau,(\delta_{H_{\bar{Q}}}\delta_{\bar{Q}}^{-1/2})\otimes\xi|_{H_{\bar{Q}}}).$$

By Proposition 3.5, $\delta_{\bar{Q}}|_{H_{\bar{Q}}} = \delta_{H_{\bar{Q}}}$. Hence we have

$$\dim(\mathrm{Hom}_{H(F)}(V_1,\xi)) = \dim(\mathrm{Hom}_{H_{\bar{Q}}(F)}(\tau,\delta_{H_{\bar{Q}}}^{1/2}\otimes\xi|_{H_{\bar{Q}}})) = m(\tau)$$

where the last equality is the definition of $m(\tau)$. □

The following proposition will be proved in the next subsection.

Proposition A.2. *With the notation above, for all $2\leq i\leq k$, we have*

$$\mathrm{Hom}_{H(F)}(V_i/V_{i-1},\omega) = \mathrm{Hom}_{Q_i(F)}(\tau,(\delta_{Q_i}\delta_{\bar{Q}}^{-1/2})\otimes{}^{\gamma_i}\xi) = \{0\}.$$

In other words, all the non-open orbits are not distinguished.

Combine Proposition A.1 and A.2 with the inequality (A.3), we have

$$m(\pi)\leq \dim(\mathrm{Hom}_{H(F)}(V_1,\xi)) = m(\tau).$$

This proves (A.1). Hence it remains to prove Proposition A.2.

A.2. **The proof of Proposition A.2.** In this subsection, we are going to prove Proposition (A.2). In other words, we need to show that the representation τ of $\bar{Q}(F)$ is not $(Q_i,(\delta_{Q_i}\delta_{\bar{Q}}^{-1/2})\otimes{}^{\gamma_i}\xi)$-distinguished for $2\leq i\leq k$ (i.e. all the non-open orbits of $\bar{Q}(F)\backslash G(F)/H(F)$ are not distinguished). We will only study the quasi-split case. The non quasi-split case follows from a similar but easier argument (this is because there are less orbits in $\bar{Q}(F)\backslash G(F)/H(F)$ for the non quasi-split case). To simplify the computation, we will use the matrix w_{2n} instead of $J_{2n,\varepsilon}$ to define the even unitary similitude group, and we set $\mathrm{GU}_{n,n} = \mathrm{GU}(w_{2n})$ (in particular, $G(F) = \mathrm{GU}_{3,3}(F)$). Since $\bar{Q}$ is a maximal parabolic subgroup, its Levi part $M_Q(F)$ is isomorphic to $\mathrm{GL}_2(E)\times\mathrm{GU}_{1,1}(F)$, $\mathrm{GL}_1(E)\times\mathrm{GU}_{2,2}(F)$ or $\mathrm{GL}_3(E)\times\mathrm{GL}_1(F)$.

We first consider the case when $M_Q(F)\simeq\mathrm{GL}_2(E)\times\mathrm{GU}_{1,1}(F)$. In this case, we may just take $\bar{Q} = \bar{P}$ and $M_Q = M$ where $\bar{P}$ is the parabolic subgroup opposite to P and $P = MU$ is the standard parabolic subgroup of G defined in the introduction. We need to compute the double coset $\bar{P}(F)\backslash G(F)/H(F)$. By the Bruhat decomposition, the double coset $\bar{P}(F)\backslash G(F)/P(F)$ contains 5 elements $\{\bar{P}(F)v_iP(F)|\ 1\leq i\leq 5\}$ with

$$v_1=I_6,\ v_2=w_{(653421)},\ v_3=w_{(623451)},\ v_4=w_{(351624)}, v_5=w_{(321654)}.$$

Here we are using partitions to denote the Weyl elements. To be specific, $w_{(s_1,\cdots,s_n)}$ is the n-by-n matrix with entries 1 in the (s_k,k) positions $(1\leq k\leq n)$ and 0 elsewhere. Then we need to break the orbit $\bar{P}(F)v_iP(F)$ into orbits in $\bar{P}(F)\backslash G(F)/H(F)$.

For $i=1,2$, it is easy to see that $M(F)\subset P(F)\cap v_i^{-1}\bar{P}(F)v_i$. Together with the fact that $P(F)=M(F)H(F)$, we have $\bar{P}(F)v_iP(F)=\bar{P}(F)v_iH(F)$ for $i=1,2$. For $i=3$, $M(F)\cap v_i^{-1}\bar{P}(F)v_i\simeq B_{\mathrm{GL}_2}(F)\times \mathrm{GU}_{1,1}(F)$ where $B_{\mathrm{GL}_2}(F)$ is the upper triangular Borel subgroup of $\mathrm{GL}_2(E)$. Since the double coset

$$B_{\mathrm{GL}_2}(F)\times \mathrm{GU}_{1,1}(F)\backslash \mathrm{GL}_2(E)\times \mathrm{GU}_{1,1}(F)/\mathrm{GU}_{1,1}(F)^{\Delta}$$

contains two orbits that are represented by $I_2\times I_2$ and $\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)\times I_2$, we have

$$\bar{P}(F)v_3P(F)=\bar{P}(F)v_{31}H(F)\cup \bar{P}(F)v_{32}H(F)$$

with

$$v_{31}=v_3,\ v_{32}=v_3\cdot diag(\begin{pmatrix}1&0\\1&1\end{pmatrix},I_2,\begin{pmatrix}1&0\\-1&1\end{pmatrix}).$$

Similarly, since the double coset

$$B_{\mathrm{GL}_2}(F)\times \bar{B}_0(F)\backslash \mathrm{GL}_2(E)\times \mathrm{GU}_{1,1}(F)/\mathrm{GU}_{1,1}(F)^{\Delta}$$

contains three elements that are represented by $I_2\times I_2,\ I_2\times \left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)\times I_2$ where $\bar{B}_0(F)$ is the lower triangular Borel subgroup of $\mathrm{GU}_{1,1}(F)$, we have

$$\bar{P}(F)v_4P(F)=\bar{P}(F)v_{41}H(F)\cup \bar{P}(F)v_{42}H(F)\cup \bar{P}(F)v_{43}H(F)$$
$$\bar{P}(F)v_5P(F)=\bar{P}(F)v_{51}H(F)\cup \bar{P}(F)v_{52}H(F)\cup \bar{P}(F)v_{53}H(F)$$

with

$$v_{41}=v_4,\ v_{42}=v_4\cdot diag(\begin{pmatrix}1&1\\0&1\end{pmatrix},I_2,\begin{pmatrix}1&-1\\0&1\end{pmatrix}),\ v_{43}=v_4\cdot w_{(124356)},$$
$$v_{51}=v_5,\ v_{52}=v_5\cdot diag(\begin{pmatrix}1&0\\1&1\end{pmatrix},I_2,\begin{pmatrix}1&0\\-1&1\end{pmatrix}),\ v_{53}=v_5\cdot w_{(124356)}.$$

To summarize, the double coset $\bar{P}(F)\backslash G(F)/H(F)$ contains 10 orbits $\{\bar{P}(F)\gamma_iH(F)|\ 1\leq i\leq 10\}$ with

$$\gamma_1=I_6,\ \gamma_2=w_{(653421)},\ \gamma_3=w_{(623451)},\ \gamma_4=w_{(623451)}\cdot diag(\begin{pmatrix}1&0\\1&1\end{pmatrix},I_2,\begin{pmatrix}1&0\\-1&1\end{pmatrix}),$$
$$\gamma_5=w_{(351624)},\ \gamma_6=w_{(351624)}\cdot diag(\begin{pmatrix}1&1\\0&1\end{pmatrix},I_2,\begin{pmatrix}1&-1\\0&1\end{pmatrix}),\ \gamma_7=w_{(351624)}\cdot w_{(124356)},$$
$$\gamma_8=w_{(321654)},\ \gamma_9=w_{(321654)}\cdot diag(\begin{pmatrix}1&0\\1&1\end{pmatrix},I_2,\begin{pmatrix}1&0\\-1&1\end{pmatrix}),\ \gamma_{10}=w_{(321654)}\cdot w_{(124356)}.$$

Now we are ready to prove Proposition A.2 for this case. For $2 \leq i \leq 10$, we need to show that

$$\mathrm{Hom}_{Q_i(F)}(\tau, (\delta_{Q_i}\delta_{\bar{Q}}^{-1/2}) \otimes {}^{\gamma_i}\xi) = \{0\} \tag{A.5}$$

with $Q_i(F) = \bar{P}(F) \cap \gamma_i H(F)\gamma_i^{-1}$. For $2 \leq i \leq 9$, one can easily show that the character $(\delta_{Q_i}\delta_{\bar{Q}}^{-1/2}) \otimes {}^{\gamma_i}\xi$ of $Q_i(F)$ is nontrivial on $\bar{U}(F) \cap \gamma_i H(F)\gamma_i^{-1}$, and then this proves (A.5) as the representation τ of $\bar{P}(F)$ is trivial on $\bar{U}(F)$. So it remains to consider the last orbit $\bar{P}(F)\gamma_{10}H(F)$. In this case, the character will be trivial on the unipotent part $\bar{U}(F) \cap \gamma_{10}H(F)\gamma_{10}^{-1}$. Hence we can get rid of this part. So it remains to consider the reductive part $(M(F), M(F) \cap \gamma_{10}H(F)\gamma_{10}^{-1})$. By an easy computation of the intersection and the character, it is enough for us to show that as a representation of $M(F) \simeq \mathrm{GL}_2(E) \times \mathrm{GU}_{1,1}(F)$, $\tau = \tau_1 \otimes \tau_2$ is not (M', ξ')-distinguished where

$$M'(F) = \{m(a,b,x,y) = \begin{pmatrix} b & 0 \\ x & b \end{pmatrix} \times \begin{pmatrix} a & y \\ 0 & b \end{pmatrix} : a, b \in E^\times, x, y \in E \text{ with } \frac{a}{b} \in F^\times \text{ and } \frac{y}{a} \in \sqrt{\alpha}F\}$$

and the character ξ' is defined to be

$$\xi'(m(a,b,x,y)) = |\frac{a}{b}|^2 \psi(\mathrm{tr}_{E/F}(\frac{x}{b})).$$

Here the $|\frac{a}{b}|^2$-part comes from the modular character and the $\psi(\mathrm{tr}_{E/F}(\frac{x}{b}))$-part comes from the character ξ. After we modulo the center, it is enough to show that τ is not (M'', ξ'')-distinguished where

$$M''(F) = \{m'(a,x,y) = \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \times \begin{pmatrix} a & y \\ 0 & 1 \end{pmatrix} : a \in F^\times, x \in E, y \in \sqrt{\alpha}F\}$$

and $\xi''(m'(a,x,y)) = |a|^2 \psi(\mathrm{tr}_{E/F}(x))$.

Let $B_0 = T_0 N_0$ be the upper triangular Borel subgroup of $\mathrm{GU}_{1,1}(F)$ and let $J_{N_0}(\tau_2)$ be the Jacquet module of τ_2 with respect to $N_0(F)$. Then in order to show $\tau = \tau_1 \otimes \tau_2$ is not (M'', ξ'')-distinguished, it is enough to show that as a representation of $\mathrm{GL}_2(E) \times T_0(F)$, the representation $\tau_1 \otimes J_{N_0}(\tau_2)$ is not (M_0, ξ_0)-distinguished where

$$M_0(F) = \{m_0(a,x) = \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \times \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} : a \in F^\times, x \in E\},$$
$$\xi_0(m_0(a,x)) = |a|^2 \psi(\mathrm{tr}_{E/F}(x)).$$

If τ_2 is supercuspidal, $J_{N_0}(\tau_2) = 0$ and hence $\tau_1 \otimes J_{N_0}(\tau_2)$ is not $(M_0(F), \xi_0)$-distinguished. When τ_2 is a discrete series, $J_{N_0}(\tau_2)$ is a one dimension representation of $T_0(F)$ with

$$|J_{N_0}(\tau_1)(\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix})| = |\frac{a}{b}|, \; a, b \in E^\times, \frac{a}{b} \in F^\times.$$

This implies that $\tau_1 \otimes J_{N_0}(\tau_2)$ is not $(M_0(F), \xi_0)$-distinguished. When τ_2 is a tempered representation but not a discrete series, $J_{N_0}(\tau_2) = \eta_1 \oplus \eta_2$ where

η_i are characters of $T_0(F)$ with

$$|\eta_i(\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix})| = |\frac{a}{b}|^{1/2},\ a, b \in E^\times, \frac{a}{b} \in F^\times.$$

This implies that $\tau_1 \otimes J_{N_0}(\tau_2)$ is not $(M_0(F), \xi_0)$-distinguished. To summarize, we have proved Proposition A.2 when $M_Q(F)$ is isomorphic to $\mathrm{GL}_2(E) \times \mathrm{GU}_{1,1}(F)$.

Then we consider the case when $M_Q(F) \simeq \mathrm{GL}_3(E) \times \mathrm{GL}_1(F)$. In this case, we may choose $\bar{Q}(F)$ to be the parabolic subgroup of $\mathrm{GU}_{3,3}(F)$ containing the lower Borel subgroup such that its Levi part is isomorphic to $\mathrm{GL}_3(F) \times \mathrm{GL}_1(F)$. By a similar argument as in the previous case, we can show that the double coset $\bar{Q}(F)\backslash G(F)/H(F)$ contains 5 orbits $\{\bar{Q}(F)\gamma_i H(F)|\ 1 \leq i \leq 5\}$ with

$$\gamma_1 = I_6,\ \gamma_2 = w_{(654321)},\ \gamma_3 = w_{(623451)},$$
$$\gamma_4 = w_{(623451)} \cdot diag(\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, I_2, \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}),\ \gamma_5 = w_{(623451)} \cdot w_{(124356)}.$$

For $2 \leq i \leq 5$, we need to show that

$$\mathrm{Hom}_{Q_i(F)}(\tau, (\delta_{Q_i}\delta_{\bar{Q}}^{-1/2}) \otimes {}^{\gamma_i}\xi) = \{0\} \tag{A.6}$$

with $Q_i(F) = \bar{Q}(F) \cap \gamma_i H(F)\gamma_i^{-1}$. The argument is similar to the previous case. For $2 \leq i \leq 4$, one can easily show that the character $(\delta_{Q_i}\delta_{\bar{Q}}^{-1/2}) \otimes {}^{\gamma_i}\xi$ of $Q_i(F)$ is nontrivial on $\bar{U}_Q(F) \cap \gamma_i H(F)\gamma_i^{-1}$, and this proves (A.6) as the representation τ is trivial on $\bar{U}_Q(F)$. For the last orbit $\bar{Q}(F)\gamma_5 H(F)$, the character will be trivial on the unipotent part $\bar{U}_Q(F) \cap \gamma_5 H(F)\gamma_5^{-1}$. Hence we can get rid of this part. So it remains to consider the reductive part $(M_Q(F), M_Q(F) \cap \gamma_5 H(F)\gamma_5^{-1})$. By an easy computation of the intersection and the character, it is enough for us to show that as a representation of $M(F) \simeq \mathrm{GL}_3(E) \times \mathrm{GL}_1(F)$, $\tau = \tau_1 \otimes \tau_2$ is not (M', ξ')-distinguished where

$$M'(F) = \{m(a, b, x, y, z) = \begin{pmatrix} b & x & z \\ 0 & b & y \\ 0 & 0 & b \end{pmatrix} \times \begin{pmatrix} a \end{pmatrix} : a \in F^\times, b \in E^\times, x, y, z \in E\}$$

and the character ξ' is defined by

$$\xi'(m(a, b, x, y, z)) = |\frac{b\bar{b}}{a}|^{1/2}\psi(\mathrm{tr}_{E/F}(\frac{x}{b} + \frac{y}{b})).$$

Here the $|\frac{b\bar{b}}{a}|^{1/2}$-part comes from the modular character and the $\psi(\mathrm{tr}_{E/F}(\frac{x}{b} + \frac{y}{b}))$-part comes from the character ξ. But this is trivial since τ is a tempered representation (in particular, the central character of τ is unitary). This proves Proposition A.2 when $M_Q(F)$ is isomorphic to $\mathrm{GL}_3(E) \times \mathrm{GL}_1(F)$.

Finally we consider the case when $M_Q(F) \simeq \mathrm{GL}_1(E) \times \mathrm{GU}_{2,2}(F)$. In this case, let $Q'(F)$ be the parabolic subgroup of $\mathrm{GU}_{3,3}(F)$ containing the lower

Borel subgroup such that its Levi part is isomorphic to $\mathrm{GL}_1(E)\times \mathrm{GU}_{2,2}(F)$. Then we can choose $\bar{Q}(F)$ to be $\delta Q'(F)\delta^{-1}$ with

$$\delta = diag(\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, I_2, \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}).$$

By a similar argument as in the previous cases, we can show that the double coset $\bar{Q}(F)\backslash G(F)/H(F)$ contains 5 orbits $\{\bar{Q}(F)\gamma_i H(F)|\ 1\leq i\leq 5\}$ with

$$\gamma_1 = I_6,\ \gamma_2 = \delta w_{(623451)},\ \gamma_3 = \delta w_{(623451)}\delta,\ \gamma_4 = \delta w_{(321654)},\ \gamma_5 = \delta.$$

For $2\leq i\leq 5$, we need to show that

$$\mathrm{Hom}_{Q_i(F)}(\tau, (\delta_{Q_i}\delta_{\bar{Q}}^{-1/2})\otimes {}^{\gamma_i}\xi) = \{0\}. \tag{A.7}$$

The argument is similar to the previous cases. For $2\leq i\leq 4$, one can easily show that the character $(\delta_{Q_i}\delta_{\bar{Q}}^{-1/2})\otimes {}^{\gamma_i}\xi$ of $Q_i(F)$ is nontrivial on $\bar{U}_Q(F)\cap \gamma_i H(F)\gamma_i^{-1}$, and this proves (A.7) as the representation τ is trivial on $\bar{U}_Q(F)$. For the last orbit $\bar{Q}(F)\gamma_5 H(F)$, the character will be trivial on the unipotent part $\bar{U}_Q(F)\cap\gamma_5 H(F)\gamma_5^{-1}$. Hence we can get rid of this part. So it remains to consider the reductive part $(M_Q(F), M_Q(F)\cap\gamma_5 H(F)\gamma_5^{-1})$. By an easy computation of the intersection and the character, it is enough for us to show that as a representation of $M(F)\simeq \mathrm{GL}_1(E)\times \mathrm{GU}_{2,2}(F)$, $\tau=\tau_1\otimes\tau_2$ is not (M',ξ')-distinguished where

$$M'(F) = \{m(a,b,x,X) = \begin{pmatrix} a \end{pmatrix} \times \begin{pmatrix} b & x & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & a & -\frac{a\bar{x}}{b} \\ 0 & 0 & 0 & a \end{pmatrix} \begin{pmatrix} I_2 & X \\ 0 & I_2 \end{pmatrix}:$$
$$a,b\in E^{\times}, x\in E, X\in Mat_{2\times 2}(E) \text{ with } \frac{a}{b}\in F^{\times}, w_2\bar{X}w_2 + {}^tX = 0\},$$

and the character ξ' is defined by

$$\xi'(m(a,b,x,X)) = |\frac{b}{a}|^{3/2}\psi(\mathrm{tr}_{E/F}(\frac{x}{b})).$$

Here the $|\frac{b}{a}|^{3/2}$-part comes from the modular character and the $\psi(\mathrm{tr}_{E/F}(\frac{x}{b}))$-part comes from the character ξ. Let $Q_0 = L_0N_0$ be the standard parabolic subgroup of $\mathrm{GU}_{2,2}(F)$ with $L_0(F)\simeq \mathrm{GL}_2(E)\times \mathrm{GL}_1(F)$ and let $J_{N_0}(\tau_2)$ be the Jacquet model of τ_2 which is a representation of $L_0(F) = \{diag(\lambda g, w_2{}^t\bar{g}^{-1}w_2)\colon g\in \mathrm{GL}_2(E), \lambda\in F^{\times}\}$. Then in order to show $\tau = \tau_1\otimes\tau_2$ is not (M',ξ')-distinguished, it is enough to show that as a representation of $\mathrm{GL}_1(E)\times L_0(F)$, the representation $\tau_1\otimes J_{N_0}(\tau_2)$ is not $(M_0(F),\xi_0)$-distinguished where

$$M_0(F) = \{m_0(a,b,x) = \begin{pmatrix} a \end{pmatrix}\times diag(\begin{pmatrix} b & x \\ 0 & b \end{pmatrix}, \begin{pmatrix} a & -\frac{a\bar{x}}{b} \\ 0 & a \end{pmatrix})\colon a,b\in E^{\times}, x\in E \text{ with } \frac{a}{b}\in F^{\times}\},$$

$$\xi_0(m_0(a,b,x)) = |\frac{b}{a}|^{3/2}\psi(\mathrm{tr}_{E/F}(\frac{x}{b})).$$

We decompose $J_{N_0}(\tau_2)$ as $\oplus_{i=1}^{k}\tau_{2,i}$ based on the central characters (i.e. the center of $L_0(F)$ acts by scalar on $\tau_{2,i}$ and the central characters of $\tau_{2,i}$ and $\tau_{2,j}$ are different for any $i\neq j$). We use χ_i to denote the central character of $\tau_{2,i}$. Since τ_2 is tempered, by Proposition III.2.2 of [14], all the characters χ_i are "positive" than the square root of the modular character. To be specific, for all $1\leq i\leq k$, we have

$$|\chi_i(diag(\begin{pmatrix} b & 0\\ 0 & b\end{pmatrix}\times\begin{pmatrix}\bar{b}^{-1} & 0\\ 0 & \bar{b}^{-1}\end{pmatrix}))|=|b\bar{b}|^{s_i},\ b\in E^{\times}$$

for some $s_i\in\mathbb{R}$ with $s_i\geq 2$. On the mean time, we have

$$\xi_0((\bar{b}^{-1})\times diag(\begin{pmatrix} b & 0\\ 0 & b\end{pmatrix}\times\begin{pmatrix}\bar{b}^{-1} & 0\\ 0 & \bar{b}^{-1}\end{pmatrix}))=|b\bar{b}|^{3/2}.$$

This implies that the representation $\tau_1\otimes J_{N_0}(\tau_2)$ is not $(M_0(F),\xi_0)$-distinguished. This proves Proposition A.2 when $M_Q(F)$ is isomorphic to $\mathrm{GL}_1(E)\times\mathrm{GU}_{2,2}(F)$. Now the proof of Proposition A.2 is complete.

References

[1] R. Beuzart-Plessis, *La conjecture locale de Gross-Prasad pour les representations tempérées des groupes unitaires.* Mémoires de la SMF 149 (2016), 1-202.

[2] R. Beuzart-Plessis, *A local trace formula for the Gan-Gross-Prasad conjecture for unitary groups: the archimedean case.* Astérisque no. 418.

[3] R. Beuzart-Plessis, *On distinguished square-integrable representations for Galois pairs and a conjecture of Prasad.* Invent. Math. 214 (2018), no. 1, 437-521.

[4] M. Brion, D. Luna, Th. Vust, *Espaces homogènes sphériques.* Inventiones mathematicae volume 84, pages 617–632 (1986)

[5] R. Beuzart-Plessis, C. Wan, *A local trace formula for the generalized Shalika model.* Duke Mathematical Journal Volume 168, Number 7 (2019), 1303–1385

[6] I. N. Bernstein and A. V. Zelevinsky, *Induced representations of reductive p-adic groups.* I, Ann. Sci. Ecole Norm. Sup. 10 (1977), no. 4, 441–472

[7] H. Jacquet, R. Langlands, Automorphic forms on GL(2) Lecture Notes 114, Springer, 1971.

[8] T. Konno, *Twisted endoscopy implies the generic packet conjecture.* Israël J. of Math. 129 (2002), 253-289

[9] T. Kaletha, A. Minguez, S. Shin, P. White, *Endoscopic Classification of Representations: Inner Forms of Unitary Groups.* Priprint, 2014.

[10] C. Mok, *Endoscopic classification of representation of quasi-split unitary groups.* Memoirs of the AMS, Volume 235 (2015), No. 1108.

[11] F.Rodier, *Modèle de Whittaker et caractères de représentations.* Noncommutative harmonic analysis, Lecture Notes in Mathematics, vol.466, eds J. Carmona, J. Dixmier and M. Vergne(Springer, Berlin, 1981), 151-171.

[12] Y. Sakellaridis, *On the unramified spectrum of spherical varieties over p-adic fields.* Compositio Mathematica 144 (2008), no. 4, 978–1016.

[13] Y. Sakellaridis, A. Venkatesh, *Periods and harmonic analysis on spherical varieties.* Astérisque No. 396 (2017).

[14] J.-L. Waldspurger, *La formule de Plancherel pour les groupes p-adiques(d'après Harish-Chandra)*, J. Inst. Math. Jussieu 2 (2003), no.2, 235-333

[15] J.-L. Waldspurger, *Une formule intégrale reliée à la conjecture locale de Gross-Prasad.* Compos. Math. 146(2010), no.5, 1180-1290.

[16] J.-L. Waldspurger, *Une formule intégrale reliée à la conjecture locale de Gross-Prasad, 2e partie : extension aux représentations tempérées.* in "Sur les conjectures de Gross et Prasad. I" Astérisque No. 346 (2012), 171-312

[17] C. Wan, *A local relative trace formula for the Ginzburg-Rallis model: the geometric side.* Memoirs of the American Mathematical Society Volume 261, Number 1263.

[18] C. Wan, *Multiplicity One Theorem for the Ginzburg-Rallis Model: the tempered case.* Trans. Amer. Math. Soc. 371 (2019), 7949-7994.

[19] C. Wan, *The local Ginzburg-Rallis model over complex field.* Pacific Journal of Mathematics 291-1 (2017), 241-256.

[20] C. Wan, *A Local Trace Formula and the Multiplicity One Theorem for the Ginzburg-Rallis Model.* PhD Thesis, 2017. Available at `https://sites.rutgers.edu/chen-wan/research/`.

[21] C. Wan, *The local Ginzburg-Rallis model for generic representations.* J. Number Theory 198 (2019), 74-123.

[22] C. Wan, *On multiplicity formula for spherical varieties.* Accepted by Journal of the European Mathematical Society

[23] C. Wan, L. Zhang, *Periods of automorphic forms associated to strongly tempered spherical varieties.* Submitted, arXiv:2102.03695.

[24] B. Xu, *On a lifting problem of L-packets.* Compos. Math. 152 (2016), no. 9, 1800–1850.

Department of Mathematics & Computer Science, Rutgers University – Newark, Newark, NJ 07102, USA
Email address: `chen.wan@rutgers.edu`

Department of Mathematics, National University of Singapore, Singapore
Email address: `matzhlei@nus.edu.sg`