QUERY-GROUNDED PREFERENCE-AWARE PERSONALIZED STICKER GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized sticker generation with intent aims to create stickers that reflect both a user's visual preferences and semantic intent. Existing reference-aware approaches rely on sticker proxies to infer semantic intent, which may dilute intent extraction when suitable visual exemplars are unavailable. We propose **Qu**ery-Grounded **P**reference-**A**ware Personalized **Sticker Gen**eration (**QuPA-StickerGen**), a framework that directly incorporates the user's textual query. First, we aggregate reusable style adapters with supervision from a vision-language model. Second, user sticker usage preferences and textual queries are mapped into a shared style space. Third, both signals are adaptively fused according to query confidence weight. This enables the generation of stickers that better align with both *how users prefer content to look* and *what they want to express*. Experiments show that our approach outperforms existing methods in both user stylistic preference and semantic alignment. Dataset and code are released.[1]

1 INTRODUCTION

Personalized sticker generation is an emerging task motivated by the extensive use of stickers in everyday communication. A previous survey of 192 users across different age and gender groups found that all respondents used stickers, averaging 7.55 stickers per day—roughly one sticker every four text messages. Users primarily used stickers to express actions, greetings, feelings, thoughts, and culture Chee et al. (2025b). These findings suggest that sticker communication is both widely adopted and frequently used, highlighting the broad relevance and practical importance of research in this area. While early research has primarily focused on sticker retrieval Zhou et al. (2025a); Er et al. (2025); Liang et al. (2025), recent work has also increasingly explored sticker generation Xu et al. (2024; 2025a); Roh et al. (2026), motivated in part by the potential to create stickers that more closely match a user's intended expression and personal preferences.

A central challenge in personalized sticker generation is therefore to satisfy two complementary requirements: capturing *what the user wants to express now* while preserving *what the user tends to like*. Existing approaches have largely addressed these requirements by conditioning generation on the user's historical stickers together with a reference sticker. Historical stickers provide a source of user-specific visual preferences, while the retrieved reference serves as a proxy for the semantic intent of the current request. However, beyond visual artifact, users semantic intent is crucial in sticker usage, which is intuitively expressed by a textual query. This raises a fundamental question: *can the user's textual query directly be leveraged for sticker generation as a representation of their current intent?*

To investigate this question, we conduct a controlled study on three signals with different setting combinations: user sticker usage history, user textual query, and a retrieved sticker reference. Across two multimodal language-model pipelines, we find that the textual query provides a strong and complementary signal for generation, while the retrieved reference does not consistently provide the same benefit when the query is available. These findings suggest that the query can serve as a direct representation of current semantic intent, without requiring an intermediate sticker retrieval step. This leads us to formulate personalized sticker generation as a joint modeling problem of *current intent* and *persistent user preference*: the former is expressed by the textual query and the latter is captured by the user sticker usage history.

[1] https://anonymous.4open.science/r/qupa-stickergen-0234/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Based on this problem formulation, we propose a lightweight QuPA-StickerGen (**Q**uery-**G**rounded **P**reference-**A**ware Personalized **Sticker Gen**eration) framework. First, we constructed global sticker style clusters based on the all users' historical stickers. Second, when a user's query comes, we select top style adapters according to query confidence weight, and then compose them into user specific style vector. During this process, no further training is needed.

Our contributions are summarized as follows:

- We formulate personalized sticker generation as a joint modeling problem of *current semantic intent* and *persistent user preference*, represented by the textual query and the user's historical stickers, respectively. Through an empirical study, we demonstrate the value of using the textual query as a direct semantic signal rather than relying on a retrieved reference sticker as a proxy for intent.
- We propose a lightweight framework that maps both query text and history preferences into a shared style space and adaptively fuses them according to query confidence. This results in reusable global style centroids, and enables user-specific style vector generation by a simple select-and-compose adapter without further training.
- We evaluate the proposed approach across multiple generation backbones and datasets, comparing against retrieval- and generation-based baselines as well as recent multimodal LLM-based approaches. Results show that jointly modeling query intent and historical preference improves the semantic alignment and personalization of generated stickers, demonstrating the effectiveness of the proposed formulation. Furthermore, the unified dataset introduced in this work can be reused by the community to support further research in personalized sticker generation and related directions.

2 Related Work

2.1 Personalized Image Generation

Recent advances in text-to-image generation have enabled models to synthesize images conditioned on user-specific visual concepts and preferences. Early personalization methods adapt pretrained diffusion models to a small set of user-provided images through parameter-efficient or lightweight optimization. DreamBooth (Ruiz et al., 2023) fine-tunes a diffusion model to bind a unique identifier to a personalized concept, while Textual Inversion (Gal et al., 2022) represents a user-provided concept as a learnable token in the text embedding space. Custom Diffusion (Kumari et al., 2023) further reduces the number of optimized parameters and supports composition of multiple personalized concepts. More recent approaches reduce the need for per-concept optimization by conditioning pretrained generators directly on reference images, such as IP-Adapter (Ye et al., 2023), or by learning efficient image encoders and adapters for personalization without test-time fine-tuning, such as InstantBooth (Shi et al., 2024).

While these methods demonstrate the feasibility of incorporating user-specific visual information into image generation, they primarily study general image synthesis. Sticker generation presents a distinct challenge: as compact visual expressions intended for communication, stickers must simultaneously convey the user's intended meaning and reflect their preferred visual style. This requires disentangling *what the user wants to express* from *how the user prefers it to be expressed*, motivating a formulation specifically tailored to personalized sticker generation.

2.2 Sticker Understanding, Retrieval, and Generation

Research on stickers spans understanding, retrieval, and generation, but remains relatively limited compared to the broader general images' literature. Sticker understanding studies their communicative roles, from emotional and contextual interpretation (Laddha et al., 2020) to multimodal recognition (Wang et al., 2024; Zheng et al., 2025) and textual semantics (Chee et al., 2025b).

Sticker retrieval includes sticker response selection, which matches stickers to conversational context (Gao et al., 2020; 2021; Chen et al., 2024; Xia et al.), and personalized sticker retrieval, which incorporates user history and multimodal representations (Zhou et al., 2025b; Er et al., 2025). De-

spite their effectiveness, retrieval methods remain constrained by the available sticker corpus and cannot synthesize a suitable sticker when no existing one matches the user's intended expression.

Recent work therefore explores personalized sticker generation. PMG (Shen et al., 2024), Pigeon (Xu et al., 2025a), and DRC (Xu et al., 2025b) incorporate user preferences into sticker or image generation, but typically rely on a reference sticker to provide semantic guidance. This couples the desired content with an existing visual exemplar, which may not adequately capture the user's intended expression. In contrast, we directly use the textual query to specify semantic intent and sticker history to capture personal preference, enabling personalized generation without requiring a semantic reference sticker.

2.3 STICKER DATASET

A growing number of publicly available sticker datasets support research on sticker understanding and retrieval, but their key signals remain fragmented. Datasets such as ChineseB2B (Zhaoolee, 2024), MCDSCS (Shi & Kong, 2024), STICKERCONV (Zhang, 2024), MultiChat (Wang et al., 2025), MOD (Fei et al., 2021), and TGIF (Li et al., 2016) provide sticker or conversational data but lack user history and query signals. Conversely, StickerQueries (Chee et al., 2025b) provides only sticker-textual query pairs but lacks user sticker usage histories. U-Sticker (Chee et al., 2025a) provides large-scale sticker conversations but does not provide sticker-textual query pairs. SRS (Gao et al., 2021; 2020), PERSRS, and WXChallenge (WeChat, 2023) provide both necessary information but are not publicly available.

Thus, existing datasets do not jointly provide publicly accessible sticker-textual queries and user sticker usage histories needed for personalized sticker generation, limiting further research.

3 UNIFYING MULTIPLE STICKER RESOURCES INTO ONE DATASET

To address the lack of dataset detailed in Section 2.3, we unify data from multiple sources into a consolidated, reusable resource for the community to develop, evaluate, and advance research on personalized sticker generation.

Specifically, we construct a user-intent sticker dataset by combining *U-Sticker* Chee et al. (2025a), a large-scale user conversation dataset containing sticker usage histories, and *StickerQueries* Chee et al. (2025b), a human-annotated dataset with search queries paired with ground-truth target stickers. The datasets are aligned via their shared sticker identifiers. This results in a multimodal dataset consisting of three components: (1) user-specific sticker usage histories, capturing long-term stylistic preferences, (2) natural language textual search queries that reflect user intent, and (3) ground-truth target stickers. This formulation jointly models personalization (via user sticker usage) and intent (via textual queries) for sticker generation. We call the subsequent dataset, STICKERGEN.

Table 1: Statistics of STICKERGEN, constructed by combining U-Sticker and StickerQueries.

# Stickers	# Searches	# Users	Avg. # History Stickers /Search	Avg. # Searches/User
31,324	8,131	925	88.42	8.79

Table 1 summarizes STICKERGEN statistics. We split the dataset by users into training, validation, and test sets with a 70/15/15 ratio, respectively, yielding 647/139/139 users. The test set contains 5,825 stickers, 139 users, 1,761 searches, and an average of 12.67 searches per user. Notably, users exhibit long interaction histories, providing rich signals for learning personalized style representations. We illustrate representative samples of the dataset in Figure 3. We provide samples of STICKERGEN in Appendix A.1. Then, STICKERGEN dataset supports a new experimental unit, (user history, query, target sticker), and a reusable resource for future research. All subsequent preliminary studies and experiments in this work are conducted on this dataset.

4 UNDERSTANDING THREE TYPES OF SIGNALS IN STICKER GENERATION

The personalized sticker generation task requires simultaneously capturing a user's established visual preferences and satisfying the semantic intent of a new request. While user history provides a natural source of personalization, it remains unclear whether the textual query alone can provide sufficient semantic guidance, or whether an explicit reference sticker is necessary to bridge the gap between the requested intent and the generated image. To disentangle these factors, we conduct an empirical study that compares three input settings: *History + Query*, *History + Reference*, and *History + Query + Reference*.

Experimental Setup. We construct a two-stage pipeline in which a multimodal understanding model converts the user's history and condition-specific query/reference signals into a generation prompt, followed by an image-generation model. We provide up to 20 historical stickers in temporal order and retrieve a non-ground-truth sticker as a reference using the textual query and BM25 Robertson et al. (2009). We evaluate two model families: GPT-5.4 Nano + GPT-Image-1 and GLM-4.6V + CogView2, with 100 instances sampled per setting. Complete prompts and additional experimental details are provided in Appendix B.1 and Appendix B.2.

Table 2: Comparison of different settings across two LLM families. Best and second-best results within each dataset are marked in **bold** and underline, respectively. Significance markers indicate statistical significance ($^{*}\ p < 0.05$, $^{**}\ p < 0.01$).

Style Alignment

Setting	CS ↑	CIS ↑	DIS ↑	LPIPS ↓	CSD_{hist} ↑	CSD_{gt} ↑	Avg. ↑
			GPT-5.4-Nano + Image-1				
His. + Ref.	0.2040**	0.5964**	0.4953**	0.6637*	0.4089**	0.3836**	0.4041**
His. + Ref. + Query	0.2088**	0.6133**	0.5101**	0.6636**	0.4250**	0.4273**	0.4201**
His. + Query	**0.2163**	**0.6195**	**0.5201**	**0.6511**	**0.4408**	**0.4347**	**0.4300**
			GLM-4.6V + CogView-2				
His. + Ref.	0.1978**	0.6234**	0.4902**	0.6761**	0.4249*	0.4021**	0.4104**
His. + Ref. + Query	0.1990**	0.6275**	0.4815**	0.6746**	0.4264**	0.4212**	0.4135**
His. + Query	**0.2111**	**0.6311**	**0.4979**	**0.6635**	**0.4440**	**0.4492**	**0.4283**

Semantic Alignment + Fidelity

Setting	CS ↑	CIS ↑	DIS ↑	CLIP-query ↑	Avg. ↑	FID ↓
			GPT-5.4-Nano + Image-1			
His. + Ref.	0.2008**	0.5911**	0.4717**	0.1683**	0.3579**	192.06
His. + Ref. + Query	0.2095**	0.6378**	0.4991**	**0.1879****	0.3836**	**190.63**
His. + Query	**0.2154**	**0.6398**	**0.5099**	0.1876	**0.3882**	193.93
			GLM-4.6V + CogView-2			
His. + Ref.	0.1909**	0.6176**	0.4665*	0.1670**	0.3605**	171.57
His. + Ref. + Query	0.2008**	0.6327**	0.4832**	0.1831**	0.3750**	170.12
His. + Query	**0.2106**	**0.6500**	**0.4951**	**0.1845**	**0.3850**	**167.17**

Evaluation Metrics. Following previous work Xu et al. (2025a), we evaluate generated stickers from three perspectives: *style alignment*, *semantic alignment*, and *generation fidelity*. For style alignment, we compare the user's sticker history with the generated sticker using CS, CIS, DIS, LPIPS Zhang et al. (2018), and CSD_{hist} Liao et al. (2024). We additionally report CSD_{gt} to measure the stylistic similarity between the generated sticker and the ground-truth sticker. For semantic alignment, we measure CS, CIS, and DIS between the generated sticker and the ground truth target sticker, and additionally introduce CLIP-query Radford et al. (2021) to directly measure the semantic alignment between the generated sticker and the sticker textual query. For fidelity, we report FID Heusel et al. (2017). Detailed metric definitions are provided in Appendix B.3.

Results. Table 2 shows a consistent pattern across both model families. *History + Query* achieves the strongest performance across nearly all style and semantic alignment metrics for both GPT and GLM, substantially outperforming *History + Reference*. This indicates that the textual query itself provides a strong semantic grounding signal while allowing the generation process to remain conditioned on the user's historical visual preferences. Interestingly, adding the query to the reference-based setting does not further improve performance. Instead, *History + Reference + Query* consistently underperforms *History + Query*. This suggests that the retrieved reference may introduce noise rather than additional visual characteristics that improve alignment with the user's intended sticker. In contrast, directly conditioning on the query avoids this additional visual constraint while retaining the user's history as the primary source of personalized style information.

These results provide empirical support for the *History + Query* formulation as a simple and effective paradigm for personalized sticker generation. More importantly, the consistent trend across two different LLM families suggests that this observation is not specific to a particular model, motivating our query-grounded preference-aware personalization sticker generation framework.

5 QuPA-StickerGen Framework

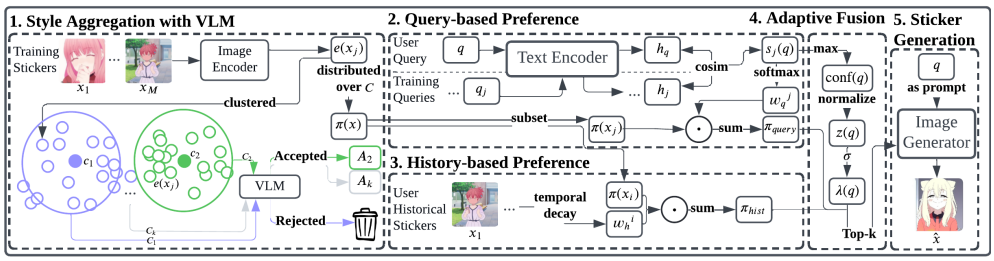

Figure 1: Overview of QuPA-StickerGen framework.

Motivated by the preceding empirical results showing that *History + Query* is an effective conditioning setting for personalized sticker generation, we propose QuPA-StickerGen, a backbone-agnostic preference modeling framework that represents both personalization and textual queries as distributions over reusable visual styles. These distributions are adaptively fused to select and compose a small set of style adapters for personalized generation. As illustrated in Figure 1, the framework consists of five components: (1) style aggregation with vision-language model, (2) query-based preference, (3) history-based preference, (4) adaptive fusion, and (5) sticker generation.

1) Style Aggregation with Vision-Language Model. We construct a reusable style space from the training stickers $\mathcal{X}_{\text{train}} = \{x_j\}_{j=1}^{M}$. Let $e\colon \mathcal{X} \to \mathbb{R}^d$ denote the CSD style encoder. Each sticker is represented by its ℓ_2-normalized embedding $e(x_j)$. We cluster these embeddings into K style clusters, using GPT-4o-mini to reject candidate merges that would conflate visually distinct styles. For each cluster $\mathcal{C}_k$, we train one LoRA adapter A_k and compute the cluster centroid $c_k = \ell_2\text{-normalize}\Big(\frac{1}{|\mathcal{C}_k|}\sum_{j\in\mathcal{C}_k} e(x_j)\Big)$. Stacking these centroids gives

$$C = [c_1, \ldots, c_K] \in \mathbb{R}^{d\times K}. \tag{1}$$

For any sticker x, its soft assignment over the style space is

$$\pi(x) = \text{softmax}\left(\frac{e(x)^\top C}{\tau_s}\right) \in \Delta^{K-1}, \tag{2}$$

where $\tau_s > 0$ is a style-temperature hyperparameter. We provide the VLM prompt in Appendix C.1.

2) Query-based Preference. Let $h\colon \mathcal{Q} \to \mathbb{R}^{d_q}$ denote the SBERT sentence encoder. We encode each training query q_j and the input query q as $h_j = h(q_j)$ and $h_q = h(q)$. Each training query q_j is paired with ground-truth sticker x_j and thus inherits its style distribution $\pi(x_j)$. The semantic similarity between the input query and each training query is

$$s_j(q) = \cos(h_q, h_j). \tag{3}$$

The corresponding query-retrieval weight is

$$w_j^q = \frac{\exp([s_j(q) - s_{\max}(q)]/\tau_q)}{\sum_{l=1}^{M} \exp([s_l(q) - s_{\max}(q)]/\tau_q)}, \tag{4}$$

where $s_{\max}(q) = \max_l s_l(q)$ and $\tau_q > 0$ controls the sharpness of the semantic weighting. The query-induced style preference is

$$\pi_{\text{query}} = \sum_{j=1}^{M} w_j^q \, \pi(x_j). \tag{5}$$

3) History-based Preference. Let $\mathcal{H} = \{x_i\}_{i=1}^{N}$ denote the user's historical stickers, ordered from oldest to newest. We assign greater importance to recent stickers:

$$w_i^h = \frac{\exp(-(N-i)/\tau_h)}{\sum_{l=1}^{N} \exp(-(N-l)/\tau_h)}, \tag{6}$$

where $\tau_h > 0$ controls the temporal decay rate. The history-induced style preference is

$$\pi_{\text{hist}} = \sum_{i=1}^{N} w_i^h \, \pi(x_i). \tag{7}$$

When no user history is available ($N = 0$), we fall back to the population prior $\pi_{\text{pop}} = \frac{1}{M}\sum_{j=1}^{M} \pi(x_j)$, setting $\pi_{\text{hist}} = \pi_{\text{pop}}$.

4) Adaptive Fusion. The query and history experts provide complementary signals: the query captures current intent, while the history captures persistent visual preferences. Because query reliability varies, we adaptively weight their contributions.

We measure query confidence by its strongest semantic match in the training set:

$$\text{conf}(q) = \max_j \cos(h_q, h_j). \tag{8}$$

Let μ_c and σ_c be the mean and standard deviation of this statistic on a held-out reference set. We compute the standardized confidence

$$z(q) = \frac{\text{conf}(q) - \mu_c}{\sigma_c}, \tag{9}$$

and the query mixing weight

$$\lambda(q) = \text{sigmoid}(a + b\, z(q)), \tag{10}$$

where $a, b \in \mathbb{R}$ are learned scalar parameters. A high-confidence query increases the query expert's contribution ($\lambda \to 1$); a low-confidence query shifts the balance toward user history ($\lambda \to 0$).

We fuse the two preferences via a weighted geometric mean, then renormalize:

$$\tilde{\pi}_k = \frac{\pi_{\text{hist},k}^{1-\lambda} \, \pi_{\text{query},k}^{\lambda}}{\sum_{r=1}^{K} \pi_{\text{hist},r}^{1-\lambda} \, \pi_{\text{query},r}^{\lambda}}. \tag{11}$$

We retain the K' adapters with the largest fused weights:

$$\mathcal{K}^* = \text{top-}K'(\tilde{\pi}), \tag{12}$$

and renormalize their weights over the selected set:

$$\alpha_k = \frac{\tilde{\pi}_k}{\sum_{r \in \mathcal{K}^*} \tilde{\pi}_r}, \qquad k \in \mathcal{K}^*. \tag{13}$$

5) Sticker Generation. The selected adapters $\{A_k\}_{k \in \mathcal{K}^*}$ are composed using their normalized weights $\{\alpha_k\}_{k \in \mathcal{K}^*}$; let $\mathcal{A}(\alpha)$ denote the resulting composed adapter. The personalized sticker is generated as

$$\hat{x} \sim p_{\theta,\, \mathcal{A}(\alpha)}(x \mid q), \tag{14}$$

where θ are the frozen base model parameters. All preference inference and adapter selection rely on precomputed embeddings, requiring no user-specific optimization at inference time. The preference modeling and fusion mechanism is independent of the text-to-image backbone, while the style adapters are trained for the chosen backbone.

6 EXPERIMENTS

6.1 SETTINGS

Baseline and Metrics. We compare against three categories of baselines. *Generation baselines* include DRC Xu et al. (2025c), Pigeon Xu et al. (2025a), and PMG Shen et al. (2024). *Retrieval baselines* include Global Popularity, which always selects the most popular sticker, Random, which randomly selects a sticker, and Recent History, which selects the user's most recently used sticker. *LLM-generation-pipelines* include OpenAI GPT-5.4-Nano + Image-1 and GLM-4.6V + CogView-2.For DRC, Pigeon, PMG, and the retrieval baselines, we evaluate on the full test set. For the LLM–generation pipelines, GPT-5.4-Nano + Image-1 and GLM-4.6V + CogView-2, we randomly sample 100 test instances from each dataset and compare their generated outputs against those of our framework. We use the same metrics as stated in Section 4. We provide the detailed hyperparameters in Appendix B.1 and the LLM-pipeline prompts in Appendix B.2.

Datasets and Image Generators. We evaluate on two datasets: WXCHALLENGE WeChat (2023), a sticker retrieval dataset which contains user, query, history, ground-truth sticker, and STICKERGEN, the unified sticker dataset, constructed as described in Section 3. We employ two representative text-to-image generation backbones, FLUX.1-DEV Labs (2024) and STABLE DIFFUSION XL (SDXL) Podell et al. (2023), to evaluate the effectiveness of our framework across different generation architectures.

6.2 OVERALL PERFORMANCE

Table 3: Performance comparison of our framework against baselines on STICKERGEN and WXCHALLENGE.

	Style Alignment							Semantic Alignment					Fidelity
	Gen. vs. Hist.					Gen. vs. Ref.		Gen. vs. Ref.			Gen. vs. Query		
Method	CS ↑	CIS ↑	DIS ↑	LPIPS ↓	CSD_{hist}	CSD_{gt}	Avg.	CS ↑	CIS ↑	DIS ↑	CLIP-query ↑	Avg.	FID ↓
STICKERGEN													
Generation													
Pigeon	**0.1885**	0.6465**	**0.4832****	0.6776**	0.4335**	0.4339**	0.4180	0.1837**	0.6513**	0.4890**	0.1685**	0.3731	78.62
DRC	0.1847**	0.6380**	0.4600**	0.6936	0.4093**	0.4211**	0.4033	0.1823**	0.6473**	0.4684**	0.1697**	0.3669	80.60
PMG	0.1319**	0.4667**	0.0301**	0.9188**	0.0619**	0.0495**	0.1369	0.1313**	0.4661**	0.0231**	0.1459**	0.1916	369.44
Retrieval													
Global Pop.	0.1669**	0.6044**	0.4111**	0.7070**	0.3217**	0.3347**	0.3553	0.1621**	0.6134**	0.4270**	0.1597**	0.3406	222.95
Random	0.1445**	0.5580**	0.2699**	0.7255**	0.2961**	0.3008**	0.3073	0.1423**	0.5635**	0.2727**	0.1588**	0.2843	**38.40**
Recent His.	0.1734**	**0.6794****	0.4817*	**0.5220****	**0.5176****	0.4925**	**0.4704**	0.1666**	0.6613**	0.4494**	0.1670**	0.3611	43.45
Ours	**0.1888**	0.6558	0.4740	0.6956	0.4699	**0.5576**	0.4417	**0.1954**	**0.7093**	**0.5444**	**0.1898**	**0.4097**	107.94
WXCHALLENGE													
Generation													
Pigeon	**0.1772**	0.6719**	0.2847**	0.7122**	0.3339**	0.3224**	0.3463	**0.1736**	0.6650**	0.2647**	0.1649**	0.3170	176.57
DRC	0.1693**	0.6132**	0.2323**	0.7894**	0.2783**	0.2739**	0.2963	0.1677**	0.6109**	0.2346**	0.1592**	0.2931	258.09
PMG	0.1225**	0.4941**	0.0477**	0.8241**	0.0890**	0.0822**	0.1686	0.1216**	0.4881**	0.0484**	0.1410**	0.1998	425.87
Retrieval													
Global Pop.	0.1666**	**0.7602****	0.2082**	0.8102**	0.3403**	0.3339**	0.3332	0.1671**	**0.7536****	0.2039**	0.1726**	0.3243	265.09
Random	0.1571**	0.7050**	0.2250**	0.7405**	0.3525**	0.3425**	0.3403	0.1565**	0.6954**	0.2283**	0.1687**	0.3122	106.33
Recent His.	0.1595**	0.7176**	0.2733**	**0.5857****	0.3924**	0.3435**	0.3834	0.1556**	0.6906**	0.2256**	0.1699*	0.3104	**48.28**
Ours	**0.1770**	0.7404	**0.3021**	0.7185	**0.4423**	**0.4151**	**0.3931**	**0.1760**	0.7305	**0.2756**	**0.1728**	**0.3387**	146.44

Main Results. Table 3 compares QuPA-StickerGen with generation and retrieval baselines on STICKERGEN and WXCHALLENGE. On STICKERGEN, our method achieves the strongest semantic alignment and the highest CSD_{gt}, while maintaining competitive style alignment despite generating novel stickers rather than directly retrieving from the user's history. On WXCHALLENGE, it similarly achieves the strongest aggregate style and semantic alignment. Across both datasets, these results demonstrate that our framework can transfer historical stylistic preferences while remaining aligned with the user's current query.

In addition, we compare our framework with GPT- and GLM-based LLM–generation pipelines on a randomly sampled subset of $n = 100$ queries. As shown in Table 4, our method achieves competitive performance with both pipelines, with stronger overall alignment on STICKERGEN. One possible explanation is that these pipelines represent visual preferences through an intermediate textual description, which may not preserve fine-grained visual characteristics as directly as our shared

Table 4: Performance comparison against state-of-the-art LLM–generation pipelines on a randomly sampled subset of $n = 100$. Our framework achieves competitive performance with both GPT- and GLM-based pipelines, with particularly strong style and semantic alignment on StickerGen.

Style Alignment

Method	CS ↑	CIS ↑	DIS ↑	LPIPS ↓	CSD_{hist}	CSD_{gt}	Avg.
StickerGen							
GPT-5.4-Nano+Image-1	**0.2163****	0.6195**	**0.5201**	0.6511**	0.4408**	0.4347**	0.4300*
GLM-4.6V+Cogview-2	0.2111**	**0.6311***	0.4979	**0.6635****	**0.4440****	0.4492**	0.4283**
Ours	0.1873	**0.6495**	0.5039	0.6962	**0.4826**	**0.5433**	**0.4451**
WXChallenge							
5.4-Nano+Image-1	0.1987**	0.7214**	**0.3273****	0.6845**	0.3744**	0.3770**	0.3857
GLM-4.6V+CogView-2	**0.2086****	0.6909**	0.2871	0.7107**	0.3331**	0.3278**	0.3561
Ours	0.1751	**0.7429**	0.2964	**0.7227**	**0.4472**	**0.4182**	**0.3929**

Semantic Alignment + Fidelity

Method	CS ↑	CIS ↑	DIS ↑	CLIP-query ↑	Avg.	FID ↓
StickerGen						
GPT-5.4-Nano+Image-1	**0.2154****	0.6398**	0.5099*	0.1876	0.3882*	193.93
GLM-4.6V+Cogview-2	0.2106**	0.6500**	0.4951**	0.1845	0.3850*	**167.17**
Ours	0.1921	**0.6950**	**0.5441**	**0.1910**	**0.4055**	179.60
WXChallenge						
5.4-Nano+Image-1	0.1972**	**0.7348**	**0.3182****	**0.1910****	**0.3603**	255.28
GLM-4.6V+CogView-2	**0.2085****	0.6894**	0.2689	0.1745	0.3353	263.45
Ours	0.1723	0.7275	0.2623	0.1760	0.3345	**183.10**

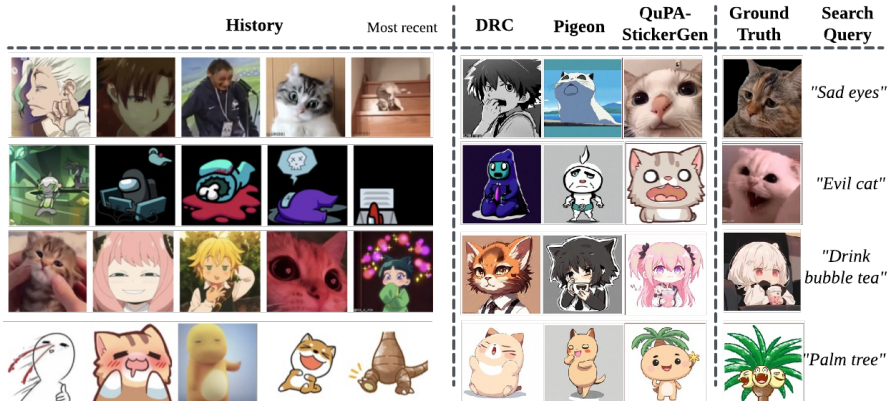

Figure 2: Qualitative comparison with generative baselines.

Table 5: Ablation study across different mechanisms, datasets and backbones.

	StickerGen						WXChallenge					
	SDXL			FLUX			SDXL			FLUX		
Model	Style	Sem.	Overall	Style	Sem.	Overall	Style	Sem.	Overall	Style	Sem.	Overall
Base	.2061	.2390	.2226	.3341	.3363	.3352	.1706	.1734	.1871	.1563	.1728	.1824
1. History Only	**.4505**	.3995	.4250	**.4222**	.3866	.4044	**.3952**	.1725	**.3679**	.3635	.1688	.3505
2. Query Only	.4306	.4067	.4187	.4111	.3904	.4008	.3876	**.1735**	.3644	.3595	**.1691**	.3472
3. History + Query	.4417	**.4097**	**.4258**	.4178	**.3929**	**.4054**	.3931	.1728	.3659	**.3658**	.1690	**.3516**

visual style representation. In contrast, our framework represents user preferences directly in the visual space, allowing such stylistic information to be retained. In addition, we provide qualitative studies between QuPA-StickerGen, Pigeon and DRC, shown in Figure 2. More qualitative examples can be found in Appendix D.

Human Evaluation. We conduct a human evaluation on 150 randomly sampled cases with 40 raters, evaluating both style and query alignment. Each case was presented three times, and the resulting ratings were aggregated across presentations. For style alignment, raters were provided with a series of stickers from the user's usage history and asked to score each generated sticker on a four-point scale ranging from 1 (inconsistent) to 4 (highly consistent), with 2 and 3 corresponding to fair and consistent, respectively. Our method achieved a mean score of 2.19 (95% CI: [2.10, 2.28]), compared with 2.08 (95% CI: [1.99, 2.17]) for Pigeon and 2.06 (95% CI: [1.97, 2.15]) for DRC. For query alignment, raters performed a forced-choice comparison among the three methods. Our method was preferred in 45.3% of cases (95% CI: [40.1%, 50.6%]), compared with 27.8% for DRC and 26.0% for Pigeon. The preference for our method was significantly above the 33.3% chance level. Overall, the human evaluation shows that our method achieves a higher mean style-alignment score than both baselines, while demonstrating a clear preference advantage for query alignment, which is aligned with Table 3. The raters reported diverse language backgrounds, including English, Chinese, Russian, German, Turkish, and Oromo.

6.3 Ablation Study and Image Generator Analysis

The ablation results in Table 5 demonstrate the complementary roles of user history and query information. History-only conditioning substantially improves style alignment, while query-only conditioning generally provides stronger semantic alignment. Combining both signals achieves the best Overall score in most settings, indicating that history captures *how the user prefers content to look*, while the query specifies *what the user wants to express*. These trends are consistent across both SDXL and FLUX and across StickerGen and WXChallenge.

6.4 Preference Component Sensitivity

We further examine the number of preference components used by the LoRA-based personalization module. As shown in Table 6, reducing the number of components from the full setting to $k = 1$ consistently degrades both style and semantic alignment, resulting in an Overall score of 0.4119 compared with 0.4258. Increasing to $k = 2$ recovers most of the performance, reaching 0.4208. The relatively small gap between $k = 2$ and the full setting suggests that the benefit of additional components saturates quickly. Overall, results indicate that modeling multiple preference is beneficial, while performance becomes relatively stable once a small number of components is retained.

Table 6: Sensitivity to the number of preference components k on StickerGen with SDXL.

k	Style ↑	Semantic ↑	Overall ↑
1	.4266	.3972	.4119
2	.4366	.4049	.4208
Full	**.4417**	**.4097**	**.4258**

7 Conclusion

In this work, we formulate personalized sticker generation as jointly modeling current semantic intent from a textual query and persistent stylistic preference from user history, and empirically show the advantage of using the query as a direct semantic signal. We introduce StickerGen, a reusable dataset that enables the experimental unit $(\text{user history}, \text{query}, \text{target sticker})$, and propose **QuPA-StickerGen**, a lightweight, generator-agnostic framework that adaptively fuses these two signals. Experiments across datasets and generation backbones, together with human evaluation, demonstrate improved semantic and personalized alignment over existing baselines, while **StickerGen** provides a reusable resource for future research on personalized visual generation.

AI USE STATEMENT

In this work, we used generative AI tools for polishing and editing the writing, as well as for writing code used in experiments and the human evaluation platform.

We have not used generative AI tools for retrieval or discovery of research literature, to draft large sections of the paper, or to prove mathematical claims. Other required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for tasks such as debugging and improving experimental code, refining wording and presentation, and providing suggestions during implementation. We have reviewed all AI-assisted work. AI-generated code was manually inspected, tested, and verified for correctness before use in the experiments and human evaluation platform. All AI-assisted text was reviewed and edited by the authors for accuracy, clarity, and consistency with the work. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Heng Er Metilda Chee, Jiayin Wang, Zhiqiang Guo, Weizhi Ma, Qinglang Guo, and Min Zhang. A 106k multi-topic multilingual conversational user dataset with emoticons, 2025a. URL https://arxiv.org/abs/2502.19108.

Heng Er Metilda Chee, Jiayin Wang, Zhiqiang Guo, Weizhi Ma, and Min Zhang. Small stickers, big meanings: A multilingual sticker semantic understanding dataset with a gamified approach, 2025b. URL https://arxiv.org/abs/2506.01668.

Jiali Chen, Yi Cai, Ruohang Xu, Jiexin Wang, Jiayuan Xie, and Qing Li. Deconfounded emotion guidance sticker selection with causal inference. In *ACM Multimedia 2024*, 2024.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: mastering text-to-image generation via transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Chee Heng Metilda Er, Jiayin Wang, Zhiqiang Guo, Chuang Ying, Jingyuan Zhang, Gang Song, Dawei Zhang, and Haifeng Wu. Persrv: Personalized sticker retrieval with vision-language model. In *Proc. of the Web Conf. (WWW)*, 2025.

Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark. *arXiv preprint arXiv:2109.01839*, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Shen Gao, Xiuying Chen, Chang Liu, Li Liu, Dongyan Zhao, and Rui Yan. Learning to respond with stickers: A framework of unifying multi-modality in multi-turn dialog. In *Proceedings of The Web Conference 2020*, WWW '20, pp. 1138–1148, 2020. ISBN 9781450370233. doi: 10.1145/3366423.3380191. URL https://doi.org/10.1145/3366423.3380191.

Shen Gao, Xiuying Chen, Li Liu, Dongyan Zhao, and Rui Yan. Learning to respond with your favorite stickers: A framework of unifying multi-modality and user preference in multi-turn dialog. *ACM Trans. Inf. Syst.*, 39(2), February 2021. ISSN 1046-8188. doi: 10.1145/3429980. URL https://doi.org/10.1145/3429980.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Black Forest Labs. Flux: A foundation model for image generation. *arXiv preprint arXiv:2408.12220*, 2024.

Abhishek Laddha, Mohamed Hanoosh, Debdoot Mukherjee, Parth Patwa, and Ankur Narang. Understanding chat messages for sticker recommendation in messaging apps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13156–13163, 2020.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description, 2016. URL https://arxiv.org/abs/1604.02748.

Bin Liang, Bingbing Wang, Zhixin Bai, Qiwei Lang, Mingwei Sun, Kaiheng Hou, Lanjun Zhou, Ruifeng Xu, and Kam-Fai Wong. Reply with sticker: New dataset and model for sticker retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3099–3110, 2025. doi: 10.1109/TASLPRO.2025.3587415.

Yuxi Liao et al. Measuring style similarity in diffusion models. In *arXiv preprint arXiv:2404.01292*, 2024.

OpenAI. Gpt-5.4 nano. https://developers.openai.com/api/docs/models/gpt-5.4-nano, 2026. Accessed: 2026-09-22.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Changhyun Roh, Yonghyun Jeong, Jonghyun Lee, Chanho Eom, and Jihyong Oh. Seal: Semantic-aware single-image sticker personalization with a large-scale sticker-tag dataset, 2026. URL https://arxiv.org/abs/2604.26883.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. of CVPR*, 2023.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 3833–3843, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645633. URL https://doi.org/10.1145/3589334.3645633.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8543–8552, 2024.

Yuanchen Shi and Fang Kong. Integrating stickers into multimodal dialogue summarization: A novel dataset and approach for enhancing social media interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9525–9534, 2024.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Jiale Tang, Ji Qi, Junhui Ji, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Bingbing Wang, Shijue Huang, Bin Liang, Geng Tu, Min Yang, and Ruifeng Xu. What do they "meme"? a metaphor-aware multi-modal multi-task framework for fine-grained meme understanding. *Knowledge-Based Systems*, 294:111778, 2024. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2024.111778. URL https://www.sciencedirect.com/science/article/pii/S095070512400412X.

Bingbing Wang, Yiming Du, Bin Liang, Zhixin Bai, Min Yang, Baojun Wang, Kam-Fai Wong, and Ruifeng Xu. A new formula for sticker retrieval: Reply with stickers in multi-modal and multi-session conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25327–25335, 2025.

WeChat. Wxchallenge/algo: Wechat algorithm challenge. https://algo.weixin.qq.com/. 2023. Accessed: 2024.

Wuyou Xia, Shengzhe Liu, Qin Rong, Guoli Jia, Eunil Park, and Jufeng Yang. Perceive before respond: Improving sticker response selection by emotion distillation and hard mining. In *ACM Multimedia 2024*.

Yiyan Xu, Wenjie Wang, Yang Zhang, Tang Biao, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. *arXiv preprint arXiv:2410.14170*, 2024.

Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 264–274. ACM, April 2025a. doi: 10.1145/3696410.3714843. URL http://dx.doi.org/10.1145/3696410.3714843.

Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. Drc: Enhancing personalized image generation via disentangled representation composition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025b.

Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. Drc: Enhancing personalized image generation via disentangled representation composition, 2025c. URL https://arxiv.org/abs/2504.17349.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Yiqun Zhang. Stickerconv: Generating multimodal empathetic responses from scratch, 2024. URL https://underline.io/lecture/102721-stickerconv-generating-multimodal-empathetic-responses-from-scratch.

Zhaoolee. Chinesebqb, 2024. URL https://github.com/zhaoolee/ChineseBQB. Accessed: 2025-05-27.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. Multi-granular multimodal clue fusion for meme understanding, 2025. URL https://arxiv.org/abs/2503.12560.

Changjiang Zhou, Ruqing Zhang, Jiafeng Guo, Yu-An Liu, Fan Zhang, Ganyuan Luo, and Xueqi Cheng. A generative framework for personalized sticker retrieval, 2025a. URL https://arxiv.org/abs/2509.17749.

Changjiang Zhou, Ruqing Zhang, Jiafeng Guo, Yu-An Liu, Fan Zhang, Ganyuan Luo, and Xueqi Cheng. A generative framework for personalized sticker retrieval. In *Findings of the Conf. on Empirical Methods in NLP (EMNLP)*, 2025b.