

Zero-shot Image Personalization from Personas

Harini S I^{*1}, Somesh Singh^{*1,2,3}, Yaman K Singla¹,
David Doermann², and Rajiv Ratn Shah³

¹ Adobe Media and Data Science Research (MDSR)

² SUNY at Buffalo

³ IIT-Delhi

✉ behavior-in-the-wild@googlegroups.com



Fig. 1: Overview of our zero-shot personalization framework. User personas are constructed by ranking subreddits using attention scores (α) from a graph attention network trained on the user-subreddit bipartite graph. For each user (e.g., u_1), the attention weights ($\alpha_{11}, \alpha_{21}, \alpha_{31}$) identify subreddits (e.g., r/Nature, r/Photography, r/Brazil) most characteristic of their interests. The comments and posts are sampled by importance and summarized to generate natural language **persona verbalizations**. These verbalized personas capture **demographic attributes** (location, age, profession), **interests** (art, anime, history), and **affective traits** (joyful, nostalgic, calm), highlighted in color in the persona boxes. These personas serve as roleplay prompts for **zero-shot personalization in diffusion models**, rewriting each base prompt to integrate user preferences. Generated images reflect these personalized attributes: the *photographer from São Paulo* produces vibrant street scenes, while the *student from Kyoto* generates calm, anime-inspired visuals aligned with their persona.

Abstract. Text-to-image diffusion models are increasingly deployed in creative contexts, yet remain impersonal, optimized for aggregate aesthetics rather than individual taste. Human preferences are inherently pluralistic: one user who favors muted, nostalgic portraits may prefer vibrant compositions for street photography, while another gravitates toward dreamy, overexposed film aesthetics. Existing personalization methods require dense interaction histories or per-user fine-tuning, failing in

* Equal contribution.

cold-start settings and collapsing each user’s context-dependent preferences into a single static representation. We introduce *zero-shot image personalization from personas (ZIP-P)*, a paradigm that conditions image generation on natural-language personas (concise descriptors of a user’s identity, interests, and aesthetic sensibilities) without any user-specific data or weight updates. **ZIP-P** uses an LLM in a roleplay setting to rewrite input prompts from the perspective of a given persona, steering diffusion models toward personalized outputs. To mine personas at scale, we develop an inductive Graph Attention Network over a 23M-user Reddit interaction graph with dual contrastive objectives that align graph structure with users’ visual behavior, and verbalize learned representations into coherent natural-language personas via an MLLM. We further introduce **ZIP-Bench** the first zero-shot image personalization benchmark, pairing 1.5K users with graph-mined personas and 40K generated images. Across four benchmarks and 14 LLMs spanning five model families, persona conditioning yields consistent improvements in both zero-shot and few-shot settings, with frontier models achieving the strongest gains (13–20%). In the few-shot setting, **ZIP-P** matches or exceeds fine-tuned baselines requiring per-user adapters trained on 100+ examples. Unlike baselines that collapse preferences into a fixed style, **ZIP-P** preserves intra-user preference diversity, achieving the lowest distributional divergence from users’ true preference distributions (CMMD 0.16 vs. 0.55 for fine-tuned alternatives). IPF-normalized evaluation against global population demographics further reveals that existing methods exhibit substantial bias toward narrow subpopulations, which persona conditioning significantly mitigates. A human evaluation confirms these findings: **ZIP-P** achieves a 79% win rate over generic generation and outperforms all fine-tuned baselines (58–65% win rate) without any user-specific training. We release **ZIP-Bench** at: <https://huggingface.co/datasets/behavior-in-the-wild/zip-bench>

1 Introduction

“We must design for the way people behave, not for how we would wish them to behave.” – Don Norman

Personalization has emerged as a foundational challenge in generative modeling. State-of-the-art text-to-image diffusion models such as DALL·E [20], NanoBanana [3], Firefly [9], and Qwen-Image [29] produce images of remarkable fidelity and semantic alignment, yet their outputs remain impersonal: optimized for aggregate aesthetic preferences that reflect an implicit “average user.” In practice, the same prompt elicits vastly different expectations from different people: a photographer may desire film grain and natural light, while a digital artist may prefer vibrant palettes and geometric compositions. Even a single user expresses different visual preferences depending on context: muted tones for documentary and saturated color for social media. As generative systems increasingly mediate creativity and communication, aligning their outputs with both the variation *across* users and the variation *within* a single user’s contexts constitutes a central open problem.

In principle a generative model trained across the global population and all possible contexts could solve this. In practice, this is infeasible, collecting dense, user preference histories is expensive and inherently non-scalable, with the largest publicly available datasets accounting for $<0.001\%$ of the global population. Moreover preferences evolve over time, making static preference profiles brittle. Further, existing preference annotation and collection pipelines disproportionately sample from a narrow slice of the global population, leaving a majority of potential users unrepresented. As a result, both the models and the benchmarks that evaluate them encode a narrow band of aggregate aesthetic norms that need to be continuously maintained.

Current methods attempt to solve this challenge by modeling user preferences through shallow user feedback including pair wise preferences [22, 28], prompt histories [2, 12], or detailed comments [22]. These methods fail to personalize without interaction data (the cold-start problem) or transfer across users. Moreover, user preferences shift across contexts (see Fig. 3), as a result these methods collapse each user into a single, static representation, discarding the pluralistic and context-dependent nature of human preference (see §4.2). Consequently, personalization systems must not only learn from individual user data, but also generalize effectively to users with little or no explicit feedback.

To solve this challenge, we propose personalizing image generation through natural language persona-conditioned prompt rewriting. Personas capture rich demographic, social, cultural, and behavioral contexts which influence their visual preferences. With this we enable a new paradigm called **ZIPPP** (Zero-shot Image Personalization from Personas) solving the cold start problem. By *zero-shot*, we mean that personalization requires no user-specific training examples, preference feedback, or interaction history; the method uses only a natural-language persona description as user-side input, with no per-user model updates at any stage. This is distinct from few-shot personalization (which uses a small set of preference pairs or historical prompts) and from fine-tuning-based methods (which require per-user adapter training). Concretely, ZIPPP conditions image generation by rewriting input prompts, through an LLM roleplaying the given persona. For example, a persona describing a Florence-based historian in Fig. 1 emphasizes classical art, historical atmosphere, and narrative depth; when applied to a generic prompt such as “a cozy wooden table with an open book” the LLM appends ‘weathered materials, muted tones, and historical’ to it. This rewritten prompt then conditions a downstream diffusion model, steering generation toward images aligned with the persona’s preferences.

This paradigm offers several advantages. First, it generalizes to users for whom no prior feedback is available, addressing the cold-start problem. Second, persona-conditioned prompts are explicit and editable, providing transparency, auditability and user control. Third, the approach transfers across image generation models without retraining. Finally, LLMs can simulate diverse behaviors and preferences through personas [5] and offer a higher coverage of the global population. Concretely, our contributions are fourfold:

- **Paradigm:** We introduce *zero-shot image personalization*, a paradigm that enables image personalization without explicit image preference data. ZIPPY, an LLM-based framework, rewrites input prompts conditioned on user persona. ZIPPY achieves strong and consistent additive gains, improving results in both zero- and few-shot settings compared to state-of-the-art methods (see Table 1).
- **Method:** We develop a novel Graph Attention-based inductive framework to mine personas from large-scale social graphs that are predictive of their visual preferences illustrated in Figure 2.
- **Data:** We also introduce **ZIP-Bench**, the first benchmark for zero-shot image personalization, comprising 1.5K users, their detailed and predictive natural-language personas, and 40K generated images (see Section 3), constructed using our graph neural network. We benchmark and analyze multiple open and frontier models and show consistent improvement of 3–20%.
- **Pluralism and Demographic Alignment:** We identify existing inter- and intra-user personalization challenges in current benchmarks and methods. Existing benchmarks are largely biased towards subpopulations and methods fail to personalize across diverse contexts. We propose methods to evaluate and mitigate these challenges in Section 4.2

2 Background & Related Work

A persona typically refers to a textual representation of an individual’s consistent traits, preferences, or style. Our work treats personas as zero-shot conditioning vectors for personalization. We review two areas: (1) personalization methods for text-to-image models, (2) user behavior modeling and persona creation frameworks. We position our work at the intersection of their work.

Image Personalization Methods. Personalized image generation adapts text-to-image models to reflect individual visual preferences, spanning aesthetic styles, compositional patterns, subject matter affinities, and semantic interpretations that vary across users. Existing approaches fall into two categories: *subject-driven personalization*, which reproduces specific concepts or styles [6, 18, 21, 26], and *user-based personalization*, which models broader user taste from behavioral or interaction data. User-based approaches [2, 12, 22, 28, 31] typically learn preferences from explicit feedback (liked/disliked images), historical prompts, or prompt–image pairs. However, these methods face key limitations: (1) they require dense per-user histories (10–100+ samples) or degrade with sparse or abstract inputs, (2) depend on non-scalable, and costly fine-tuning or retrieval pipelines, (3) fail to generalize across contexts for a user, and (4) fail to personalize to unseen users. For example, ViPer [22] requires 8–20 detailed annotated images and produces context-insensitive prompts (e.g., applying “vibrant colors” universally), while Tailored Vision [2] deteriorates when conditioned on more than three previous prompts.

Our approach provides a zero-shot, language-only alternative where natural-language personas derived without explicit annotations act as interpretable con-

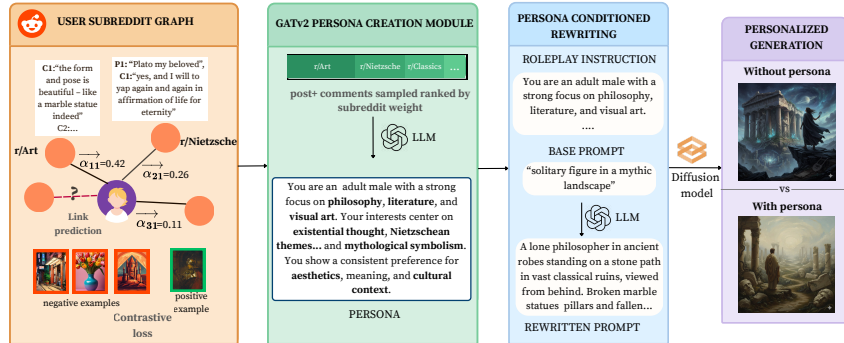


Fig. 2: Overview of methodology.

ditioning signals for personalized image generation, eliminating the need for per-user histories, feedback, or fine-tuning.

Persona Creation and Behavioral Modeling. Modeling user preferences through personas has long been central to personalization and HCI research. Recent work scales this process using large language models (LLMs). Persona Hub [7] and PersonaCraft [11] generate diverse persona corpora from web and survey data, while others [24] employ human–LLM workflows for fine-grained clustering. Role-play prompting further shows that LLMs can internalize persona traits across dialogue [16, 27], recommendation [32], search [34], and reasoning [14] tasks.

In multimodal modeling, Behavior-LLaVA [25] shows that user comments on visual content capture perceptual and emotional salience. While studied in the visual domain, this insight generalizes more broadly, as user commentary often encodes stable personal attributes such as demographics, topical interests, communication style, and aesthetic preferences. However, existing persona datasets remain either synthetic (Persona Hub) or survey-based (PersonaCraft), and behavioral traces are typically used to evaluate learned personas rather than to construct them. Our approach bridges these fields by unifying persona mining and image personalization within a single inductive framework. We extract personas from large-scale social graphs by modeling users’ network, activity, and behavior (posts and comments) and use these personas as conditioning signals for personalized image generation. Unlike prior methods that rely on dense preference histories or per-user fine-tuning, we demonstrate that natural-language personas alone can effectively guide image generation across entirely new users, contexts, and diffusion model families.

3 Methodology

Figure 2 gives an overview of our pipeline: we encode user–subreddit interactions with an attention-based graph encoder, verbalize the learned attention weights into a natural-language persona, and use that persona to condition

prompt rewriting for personalized generation. The remainder of this section details each stage, beginning with how we curate the underlying persona data before formalizing persona-driven image personalization.

3.1 Dataset Curation for Personas

Existing image personalization benchmarks and datasets suffer from two critical limitations: (1) they capture preferences from narrow, self-selected user demographics (e.g., creative hobbyists on Pick-a-Pic, expert annotators on HPD-v2), limiting generalization to diverse user populations; and (2) they lack explicit user identity representations (personas). These inhibit us from evaluating and optimizing accurately.

To address this, we construct two complementary datasets that enable large-scale persona mining and zero-shot personalization evaluation across diverse user populations.

Reddit Interactions Persona mining requires large-scale behavioral data that reveals authentic visual preferences across heterogeneous users. Unlike survey-based persona datasets (PersonaCraft [11]) or synthetically generated personas (Persona Hub [7]), we ground personas in real interaction traces from Reddit⁴, a platform with 400M+ monthly active users spanning diverse demographics and visual preferences. Reddit’s community structure provides a natural substrate for this task: users engage with topic-specific subreddits (e.g., r/AiArt) by commenting or posting, revealing their latent preferences, and in the process creating a bipartite graph of subreddits and users. We aggregate seven years of activity (2015–2022) from 23M users across 40K subreddits, capturing 682M edges of unique subreddit and user interactions. This graph structure provides strong inductive biases for persona discovery through interest neighbourhoods. Beyond Reddit, this graph-based approach can be generalized to any platform with a user-community structure. We focus on Reddit for its scale, API accessibility, and diverse visual communities spanning photography, generative art, and niche interests. In Sec. 3.1 we describe how we discover natural language personas (e.g. “student from Kyoto”) from this graph.

ZIP Bench While the Reddit graph enables persona mining at scale, benchmarking zero-shot image personalization requires users with both rich interaction histories and visual preferences or historical generations. We focus on Civitai⁵, one of the most popular open-source AI-generated content (AIGC) platforms where users create, customize, and share diffusion models and generated images. Crucially, Civitai’s public API provides detailed metadata (prompts, model versions, NSFW classifications), and many platform users maintain linked Reddit accounts, enabling us to pair graph-mined personas with real-world generation histories.

Utilizing this cross-profile linking, we identify 1.5K Civitai users active on AI art subreddits (r/civitai, r/AiArt), extracting their complete Reddit history

⁴ <https://reddit.com/>

⁵ <https://civitai.com/>

(15K posts, 183K comments from 2015-25) to mine personas from the global graph, alongside 40K generated images with prompts from their Civitai profiles. Unlike previous personalization benchmarks [2, 13], ZIP Bench uniquely combines explicit natural-language personas with generation histories, enabling the first systematic evaluation of whether and how well persona descriptions alone can reliably guide personalization without per-user fine-tuning (Sec. 4).

We now detail our two-stage pipeline to (1) learn distinct user representations from large social networks and (2) use them to create personas which are indicative of their implicit visual behavior. *persona* is a concise textual descriptor of a user’s interests, aesthetic tendencies, and engagement patterns (e.g., “anime enthusiast favoring pastel palettes and character-focused compositions”), serving as an interpretable, zero-shot conditioning signal for image personalization. Our goal is to mine such personas at scale from the Reddit interaction graph, ensuring they (1) reflect authentic behavioral patterns rather than self-reported surveys, (2) cluster users by visual preference similarity, and (3) generalize to unseen users without requiring per-user annotation.

Graph Setup: We model the Reddit interactions as a bipartite graph $\mathcal{G} = (\mathcal{U}, \mathcal{S}, \mathcal{E})$ where users $\mathcal{U}$ and subreddits $\mathcal{S}$ form distinct node types, connected by directed edges $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{S}$ representing posting and commenting activity. For each interaction $(u, s) \in \mathcal{E}$, we aggregate activity counts t_{us} (posts + comments) and compute node degrees $c_u = \sum_s t_{us}$, $c_s = \sum_u t_{us}$. Akin to social networks, we observe that Reddit-Interactions has (1) long-tailed activity distributions (12k subreddits contribute to only 0.1% of the total activity) and (2) popularity bias (popular subreddits and power-users have a very high activity). To alleviate this, we remove the bottom 12k subreddits and assign the edge weight to be log and degree normalized activity $w_{us} = \log(1 + t_{us}) / \sqrt{c_u c_s} + \varepsilon$ following PinSAGE’s [33] PMI-style normalization (Sec. E.4 shows a normal distribution on the log/log scale supporting this). The resulting graph consists of 32k subreddits, 22.8M users, and 682M edges. Additional statistics, edge weighting schemes, and ablations are in Sec. B.1. To cluster users or communities at this scale using traditional graph algorithms is infeasible; therefore, we use graph neural networks to learn these representations.

Learning User Representations It is essential to learn user representations inductively from the graph structure, enabling generalization to new users at inference without requiring retraining. Since there are only 32k subreddits and we have their detailed descriptions, we initialize the nodes with pretrained text embeddings $\mathbf{x}_s \in \mathbb{R}^{3072}$ (OpenAI text-embedding-3-large [17]) of these descriptions and user nodes with a small gaussian noise, the nodes and edges are projected into hidden states $\mathbf{h}_u, \mathbf{h}_s, \mathbf{h}_e$ respectively. We employ two GATv2 layers [1] with 4 attention heads each, aggregating information from subreddits to users via attention mechanism, the resulting user embedding $\mathbf{h}_u$ is the standard multi head attention, where the output of each attention head $\mathbf{h}_u^{(l)}$ is defined as

$$\mathbf{h}_u^{(l)} = \sum_{s \in \mathcal{N}(u)} \alpha_{us}^{(l)} \mathbf{W}_v^{(l)} \mathbf{h}_s^{(l)} \quad (1)$$

where $\alpha_{us}^{(\ell)}$ is the graph attention for the head ℓ , between the user u and its neighbors $s \in \mathcal{N}(u)$ and $\mathbf{W}, \mathbf{h}_*$ are the parameters and concatenated hidden states $\mathbf{h}_u, \mathbf{h}_s, \mathbf{h}_e$. We use the standard LeakyReLU (slope 0.2) activations, apply LayerNorm before each attention block, and L2-normalize the node embeddings $\mathbf{z}_u = \mathbf{h}'_u / \|\mathbf{h}'_u\|_2$ (and similarly $\mathbf{z}_s$ for subreddits $\mathbf{h}_s$) to maintain cosine geometry for our objectives that we define below

Dual contrastive objectives. We supervise the embedding space through the self-supervised link prediction task, we sample edges uniformly across subreddits and users, and use the InfoNCE loss to contrastively align the user embeddings $\mathbf{z}_u$ with their neighbors' embeddings $\mathbf{z}_s$ (negatives are sampled from close neighbors). Let $\mathcal{B} = \{(u_i, s_i)\}_{i=1}^B$ be a batch of edges, the objective is to minimize the loss

$$\mathcal{L}_{u,s} = \frac{1}{B} \sum_{i=1}^B \text{CE}(\mathbf{z}_{u_i}^\top [\mathbf{z}_{s_1}, \dots, \mathbf{z}_{s_B}], i), \quad (2)$$

where $\text{CE}(\cdot, i)$ is cross-entropy treating i as the positive index and all other subreddits in the batch as negatives similar to CLIP [19].

To explicitly model visual behavior signals we add a second contrastive term aligning users with the images they have posted on Reddit. For users who have ever posted images, we extract CLIP embeddings of their images, forming user-image pairs (u, img_u) and similarly, the user-image loss is:

$$\mathcal{L}_{u,\text{img}} = \frac{1}{B'} \sum_{j=1}^{B'} \text{CE}(\mathbf{z}_{u_j}^\top [\text{img}_{t_1}^t, \dots, \text{img}_{t_{B'}}^t], j) \quad (3)$$

where $B' \leq B$ is the subset of users with available image-post text. The total loss combines both objectives: $\mathcal{L} = \mathcal{L}_{u,s} + \lambda \mathcal{L}_{u,\text{img}}$, with $\lambda = 1.0$ weighting the image contrastive loss. This dual supervision ensures learned embeddings capture not only where users engage (subreddit structure) but also what visual content they share, which we hypothesize as a critical proxy for downstream persona-driven image personalization (Sec. 4). Other training details, ablations and hyperparameter settings are discussed in Sec. B.1

Persona Verbalization To construct a user's natural language persona from their subgraph, we utilize the learned attention weights α_{us} from Eq. (1) as indicators of importance. These weights represent how strongly the encoder associates each subreddit s with user u when modeling their preferences. For a given user, we aggregate attention values across their neighborhood and normalize them to obtain:

$$\tilde{\alpha}_{us} = \frac{\alpha_{us}}{\sum_{s' \in \mathcal{N}(u)} \alpha_{us'}} \quad (4)$$

This yields a probability distribution over subreddits that captures the user's attention-weighted engagement profile, highlighting which communities most characterize their interests and behaviors.

We allocate a fixed token budget $B = 4,096$ across the user's top subreddits proportional to $\tilde{\alpha}_{us}$. For each subreddit s receiving $B_s = \lfloor B \cdot \tilde{\alpha}_{us} \rfloor$ tokens, we sample the user's highest-engagement posts and comments (ranked by score,

deduplicated, filtered for removed/deleted content), extracting representative text that reveals their interests, communication style, and aesthetic preferences. This attention-guided sampling ensures personas reflect what the graph encoder learned as behaviorally salient, rather than arbitrarily frequent interactions.

The sampled content is then presented to an MLLM with a structured prompt requesting a concise, coherent persona description that captures the user’s visual interests and aesthetic tendencies. For example, a user active in r/analog, r/AnalogCommunity, and r/itookapicture with high attention weights and comments praising "grain texture" and "natural light" might yield: *"Film photography enthusiast favoring high-grain black-and-white aesthetics with natural lighting and vintage tonality."* Sec. B.2.

3.2 Image Personalization from Personas

We use the resulting natural language persona to guide zero-shot image personalization via persona-conditioned prompt rewriting. Instead of directly editing the prompt content, we use the persona as a *roleplay instruction* that defines how the model (GPT-4o) should interpret and express visual preferences. Specifically, we prepend the persona as a first-person conditioning statement (e.g., *"You are a film photography enthusiast favoring high-grain black-and-white aesthetics..."*), allowing the model to personalize generation from its own perspective. This enables implicit modulation of stylistic, compositional, and semantic attributes in line with the user’s aesthetic tendencies. We use roleplay rather than directly concatenating the persona to the prompt: direct concatenation treats persona attributes as literal scene content and underperforms even the no-persona baseline across all four benchmarks (Sec. F.3, Tab. 11). We detail the prompting format, decoding strategies, and evaluation protocols in Sec. B.3.

4 Experiments & Results

In this section, we systematically evaluate **ZIP** against state-of-the-art personalization methods. We demonstrate that natural language persona conditioning not only enables effective zero-shot alignment (Tab. 1) but also solves critical shortcomings in existing methods, namely the collapse of intra-user diversity (pluralism) (Sec. 4.2) and the bias towards average-user demographics (alignment) (Sec. 4.2). We validate our approach across four datasets for zero-shot, few-shot, and finetuning paradigms (Tab. 1). We also analyze the data efficacy of these methods (Sec. 4.2) and corroborate our automated metrics with a comprehensive human user study (Sec. 5).

4.1 Experimental Setup

Benchmarks. We evaluate our method across four complementary datasets which assess personalization using preferences and generation history. For preferences,

we utilize Flux-2-pro_t2i_human_preference⁶ from **RapidData**, which provides explicit preference pairs alongside detailed user demographics (language, country, profession, gender, and age). **MovieLens** (ML) is a movie-watching dataset covering 610 users, each with 20 to $\sim 2,600$ interactions. Following [12], we convert each movie’s metadata (title, genre, keywords, and description) into text prompts via the TMDb API; consistent with this prior work, we rely exclusively on these prompts for evaluation, as the dataset contains no images directly tied to T2I model outputs. For historical prompts, we evaluate on **ZIP-Bench** (our proposed benchmark of 1.5K users, natural-language personas, and 40K generated images mined via our approach) and **PIP** [2], which contain 300k extensive historical user generations from 3.1K users.

Baselines. We compare against five state-of-the-art personalization methods: Random Baseline **LLM** (prompt rewriting without any context) **TV** [2] (prompt rewriting via retrieval over ≤ 3 historical prompts), **DrUM** [12] (coreset sampling + per-user adapter training on 100+ prompts), **ViPer** [22] (trained VLM for preference extraction from user feedback) We compare these methods against our Zero-Shot, Few-Shot persona conditioning.

Metrics. **CLIPScore (CS).** Following prior work [2,12,28], we adopt CLIPScore as our primary metric for personalization alignment, following the exact setup of [12]: mean cosine similarity between CLIP [19] image embeddings of generated outputs and text embeddings of user target prompts or preferred images.

PIGReward (PIG). CLIP-based metrics miss nuanced stylistic and compositional preferences [15]. We additionally report PIGReward [15], a learnable reward model (Qwen2-VL-7B, DPO fine-tuned on contrastive preference pairs) that dynamically generates user-conditioned evaluation dimensions, correlating better with human judgment than CLIP-based alternatives. CS and PIG are complementary: CS measures prompt fidelity, PIG measures underlying preference alignment.

CMMD. To measure **pluralistic alignment** — whether generated images span the full spread of a user’s visual preferences rather than collapsing to a single style — we use CLIP Maximum Mean Discrepancy [10], the squared MMD between CLIP embedding distributions of reference and generated images. Unlike FID, CMMD is an unbiased estimator reliable for small per-user sample sizes; lower CMMD indicates tighter distributional agreement.

IPF-normalized score. Standard aggregate scores are biased toward WEIRD populations [8]. We apply Iterative Proportional Fitting (IPF) [4] to reweight per-user scores to match census-level marginals (country, age, gender, language, profession) from UN World Population Prospects 2024, surfacing performance disparities hidden by majority-skewed aggregates.

⁶ <https://huggingface.co/Rapidata/datasets>

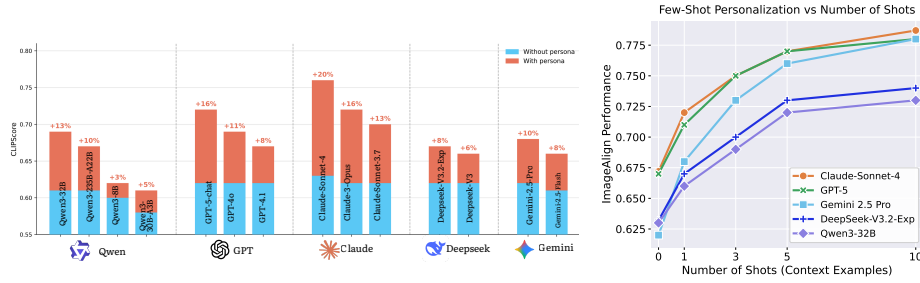


Fig. 3: Effect of persona conditioning on generation quality. (a) ClipScores by model family with persona conditioning. Stacked bars show baseline (blue) and improvement from persona conditioning (red). All models benefit from persona conditioning ranging from 3% (Qwen3-8B) to 20% (Claude-Sonnet-4). Frontier models achieve the strongest gains (13–20%), outperforming open-source alternatives (3–13%). (b) Performance increase as the number of in-context persona examples increases.

4.2 Personalization from Natural Language Persona Conditioning

We first establish that natural-language personas serve as an effective and universal personalization signal that bring consistent improvements in zero-shot and few-shot settings. Table 1 reports the unified evaluation across all four benchmarks and all baselines.

ZIP-Bench Sec. 4.2 shows the **ZIP-Bench** evaluation for ClipScore for 14 LLMs across five model families (Qwen, GPT, Claude, DeepSeek, Gemini), with and without persona conditioning (red and blue). Persona conditioning yields consistent improvements across *all* (frontier and open-source) evaluated models. Frontier models outperform open-source models due to their stronger instruction-following and persona interpretation capabilities, corroborated by recent roleplay benchmarks [5] where Claude ranks highest. Mixture-of-experts models (Qwen3-30B-A3B, Qwen3-235B-A22B) surprisingly underperform their dense counterparts despite larger parameter counts; on manual analysis we found that these models are over-personalizing. Performance of all baselines on **ZIP-Bench** is given in Table 1. Extended results across remaining metrics and hyperparameters are reported in Tab. 10 and Sec. F

Public Benchmarks. Table 1 reports CLIPScore and PIGReward across **ZIP-Bench**, PIP, RapidData, and MovieLens. For fair comparison, we use GPT-4o as the underlying LLM for all prompt-rewriting methods. Even in the **zero-shot** setting, ZIPPY already surpasses the fine-tuned DrUM baseline on three of four benchmarks, demonstrating that a compact natural-language persona captures preference signals as effectively as dense interaction histories and per-user adapter training. In the **few-shot** regime, ZIPPY (5-shot) achieves the

highest scores across all four benchmarks on both metrics. At the same shot count, ZIPPY (3-shot) consistently outperforms TV (3-shot), confirming that persona priors provide a stronger conditioning signal than retrieval over raw prompt histories alone. The only benchmark where DrUM remains competitive is PIP, which we attribute to its adapter being trained on the same dense prompt histories that define PIP users; on the remaining benchmarks, persona conditioning surpasses per-user fine-tuning without any gradient updates. To adapt our method to external benchmarks, we construct PIP user personas by embedding each user’s prompt history via a CLIP text encoder and retrieving nearest-neighbor personas from our mined set (cosine similarity ≥ 0.7 , covering 89% of PIP users); for RapidData and MovieLens, where only demographic variables and profession are available, we condition on these attributes directly. As a result, the strongest gains appear on **ZIP-Bench**, where full natural-language personas are available, followed by RapidData and MovieLens (demographics only), and finally PIP (retrieved personas). To investigate the relationship between persona richness and personalization capability in detail, we ablate summarized profiles across different token budgets in Figure 8.

Pluralistic Preference Alignment Effective personalization must capture not only *what* a user prefers on average, but also *how* their preferences vary across different contexts. A person who prefers ‘muted’ and ‘animated’ portraits could prefer ‘bright’ street photography should not get similar output for every prompt. We formalize this as *pluralistic alignment*: a personalization method should achieve high average personalization *while preserving* the distributional spread of a user’s true visual preference space. To measure the distributional alignment of target and generated images, we employ the CMMD metric, which is essentially the squared MMD of target and generated image’s clip embeddings, using with the Gaussian RBF kernel, first introduced by [10]. Lower MMD indicates closer alignment between the generated and ground-truth image distributions. We report these in Table 1. ViPer assigns the same style attributes across all contexts for the same user, making it the most susceptible to pluralistic misalignment: it achieves the worst CMMD (0.55). DrUM exhibits moderate CMMD (0.31): while its per-user adapter captures individual preferences, its coreset sampling anchors generation around a narrow slice of the user’s history. TV (0.25) benefits from retrieval diversity but still collapses when the prompt pool is small. In contrast, ZIPPY achieves the *lowest* CMMD (0.16 at 5-shot) while maintaining the *highest* alignment scores, demonstrating that persona conditioning preserves intra-user preference diversity. This occurs because personas encode multi-faceted preference priors spanning aesthetics, topics, culture, and affect, which modulate generation contextually, rather than anchoring on a fixed reference set.

Demographic Generalization Survey-based optimization and evaluation have historically suffered from sampling bias toward WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations [8]. Henrich et al. showed that up to

96% of behavioral science participants come from WEIRD societies, representing only $\sim 12\%$ of the global population, yet findings are routinely generalized as universal. This bias extends to AI systems: Santurkar et al. [23] demonstrated that language models disproportionately reflect the opinions of specific U.S. demographic groups, with alignment varying systematically across age, education, and political affiliation. To mitigate such biases, survey organizations like Pew Research apply Iterative Proportional Fitting (IPF, also known as “raking”) to reweight responses so that reported aggregates reflect true population distributions rather than the convenience sample [23]. We find that these same biases have crept into image personalization and preference alignment benchmarks. Critically, most existing benchmarks (including PIP [2], HPD [30], and Pick-a-Pic [13]) do not report individual or aggregate demographic distributions at any level, making it impossible to analyze or alleviate these biases. To address this gap, we utilize RapidData’s T2I preference datasets, a large-scale image preference dataset where each preference pair is explicitly tagged with anonymized group-level demographics (language, country, profession, age, gender). RapidData serves as a demographically transparent alternative to HPD and Pick-a-Pic, enabling the first systematic analysis of demographic bias in image personalization.

IPF normalization. Raw aggregate scores can mask systematic disparities. We extract the true demographic marginal distributions using UN World Population Prospects 2024, which provides population-level statistics for each demographic axis. We then apply Iterative Proportional Fitting (IPF) [4] to reweight per-subgroup scores, ensuring that the reported aggregate reflects a balanced population rather than the majority-skewed sample. We emphasize the importance of annotating preference datasets with anonymized demographics to ensure both privacy and equitable alignment evaluation. We show results for stratified ClipScore and PIGReward by demographic subgroup in Section F.6. Since the evaluations themselves are skewed, optimization methods that maximize aggregate population scores are most susceptible to this bias. DrUM shows the *highest* drop in scores after IPF normalization (Tab. 3), as its adapter training absorbs the distributional biases of the majority-skewed training data. In contrast, ZIPPY shows the highest alignment in all subgroups because the LLM takes the persona into its generation context. Even in the zero-shot setting, ZIPPY achieves better or comparable performance to trained or few-shot models. ViPer shows very little relative degradation after normalization ; we attribute this to the fact that ViPer’s visual preference representations are trained on synthetic visual personas that are less susceptible to demographic skew in interaction histories. On subgroup-level analysis, we observe clear stereotypical trends: DrUM shows poor alignment with Older (Age 50+), African, Russian, and South American populations and individuals in manual trades. We also observe that frontier models (Claude, GPT, Gemini) show far better demographic alignment compared to open-source alternatives (Qwen, DeepSeek), consistent with the broader instruction-following gap observed in Sec. 4.2. A detailed visu-

Train	Method	ZIP-Bench		PIP		MovieLens		RapidData		Demographic		Pluralism
		CS $\uparrow$	PIG $\uparrow$	CS $\uparrow$	PIG $\uparrow$	CS $\uparrow$	PIG $\uparrow$	CS $\uparrow$	PIG $\uparrow$	CS $\uparrow$	PIG $\uparrow$	CMMD $\downarrow$
zero shot	GPT-4o	59.1	65.3	57.7	61.4	73.7	58.5	67.3	56.0	66.9	55.1	0.49
	ZIPPY	61.8	73.4	60.2	67.3	76.4	64.2	73.9	61.5	72.8	60.9	0.42
few shot	FABRIC (1)	-	-	63.7	-	75.1	-	-	-	-	-	-
	TV (3)	62.4	72	63.1	66.5	80.1	68.7	77.4	65.3	73.8	60.1	0.25
	ZIPPY (3)	66.7	77.5	65.2	69.9	80.4	71.4	77.9	68.2	76.8	67	0.19
	ZIPPY (5)	68.5	82.8	66.3	72.1	80.9	77.9	78.3	74.0	77.4	73.0	0.16
fine tune	DrUM	65.6	74.4	66.9	68.0	72.8	66.2	70.3	63.5	64.7	56.8	0.31
	ViPer	-	-	-	-	76.1	70.5	74.1	63.1	73.7	60.8	0.55

Table 1: Unified evaluation: cross-benchmark, demographic, and pluralistic alignment. CS = CLIPScore, PIG = PIGReward, CMMD = CLIP Maximum Mean Discrepancy ($\downarrow$ is better). We report personalization benchmarks and methods for (i) personalization from historical prompts (**ZIP-Bench**, PIP) (ii) from preference pairs (MovieLens, RapidData) and across 0-shot, few-shot (FABRIC,TV), and fine-tuning baselines (DrUM, ViPer). ZIPPY achieves the highest scores across all benchmarks and the lowest CMMD, demonstrating simultaneous personalization quality, demographic equity, and pluralistic diversity preservation. To isolate the effect of persona conditioning, the zero-shot **GPT-4o** row reports the same LLM prompt-rewriting pipeline *without* any persona; the gap to ZIPPY therefore reflects the contribution of the persona alone.

alization of per-subgroup score distributions and IPF-normalized aggregates is provided in Sec. F.6.

4.3 Persona Mining Ablation

A key question is whether the graph-based persona mining pipeline actually drives the personalization gains, or whether simpler heuristics suffice. We measure two proxies: *posting accuracy* (whether the persona correctly predicts a held-out image the user posted) and *ZIP Lift* (relative CLIPScore improvement of **ZIP** over the unpersonalized baseline). TF-IDF weighting ranks subreddits by term frequency alone, causing high-traffic communities (e.g., r/AskReddit) to dominate even when they carry little signal about visual preferences. GAT addresses this by learning which subreddits are most characteristic of each user through edge-aware attention over the full bipartite graph. The ImageAlign objective further grounds embeddings in what users actually post visually, rather than only where they engage textually. As a result, GAT + ImageAlign achieves 73% posting accuracy and 17.7% ZIP Lift (Tab. 2), compared to 5.1% for TF-IDF, and ImageAlign yields consistent gains across every encoder (3–11 points in posting accuracy, 2–6% in ZIP Lift). Full hyperparams given in Sec. F.1.

5 User Study

We complement our automated metrics with a two-part human evaluation (full protocol in Sec. G): a **pairwise preference study** on perceived personalization and a **persona quality assessment** of the mined personas.

Pairwise preference. We recruit 50 participants spanning age (18–55+), gender, profession, and geography. Each answers a 10-question intake survey on demo-

Method	Post. $\uparrow$ Acc.	ZIP $\uparrow$ Lift
TF-IDF	55	5.1
TF-IDF + NMI	57	5.3
LightGCN	59	6.2
+ ImageAlign	62	9.1
GraphSAGE	63	8.5
+ ImageAlign	67	10.2
GAT	62	11.8
+ ImageAlign	73	17.7

Table 2: Graph encoder ablation. All rewrites use GPT-4o. GAT + ImageAlign achieves the best posting accuracy and ZIP Lift; green = best, blue = second best.

Method	Setting	CLIPScore			PIGReward		
		Raw	IPF	$\Delta\%$	Raw	IPF	$\Delta\%$
LLM	0-shot	67.3	66.9	-0.6	56.0	55.1	-1.6
TV	3-shot	77.4	73.8	-4.7	65.3	60.1	-8.0
DrUM	FT	70.3	64.7	-8.0	63.5	56.8	-10.6
ViPer	FT	74.1	73.7	-0.5	63.1	60.8	-3.6
ZIPPY	0-shot	73.9	72.8	-1.5	61.5	60.9	-1.0
ZIPPY	5-shot	78.3	77.4	-1.1	74.0	73.0	-1.4

Table 3: IPF degradation summary on RapidData. $\Delta\%$ is the relative change from raw to IPF-reweighted aggregate; larger negative values indicate greater degradation on underrepresented subgroups.

graphics, profession, and interests. GPT-4o converts each survey into a natural-language persona that conditions generation from 20 base prompts, and participants make *blind* A/B choices (randomized order, no method labels), yielding $50 \times 20 = 1,000$ annotations across four conditions. ZIPPY wins every condition (all $p < 0.01$, binomial test; Tab. 13): a **79%** win rate over the no-persona baseline, with an **own- vs. other-persona** control confirming the gain comes from the *correct* persona rather than generic prompt enrichment. It also beats prior methods—**56%/78%** (0-/3-shot) over TV, **58%** over DrUM, and **65%** over ViPer—despite DrUM and ViPer requiring per-user fine-tuning.

Persona quality. Three expert annotators rate 100 sampled personas ($\sim 1,000$ annotations) on a 1–5 scale for demographic accuracy, interest completeness, and coherence. Personas average **4.1/5** (4.3/3.9/4.1) with substantial agreement (Cohen’s $\kappa = 0.64$), confirming they are accurate and self-consistent. Overall, persona conditioning delivers personalization that aligns closely with human judgment.

6 Conclusion and Impact

We presented **ZIPP**, a zero-shot framework for image personalization through natural language personas. By grounding user preferences in interpretable textual descriptions, ZIPP personalizes diffusion outputs without fine-tuning or user data. Experiments and human evaluations demonstrate that persona conditioning enhances visual alignment, and that mined personas are accurate, coherent, and preferred by users. We also propose methods to evaluate and optimize the pluralism and demographic aligned of existing methods and benchmarks. All datasets are derived solely from publicly available Reddit content, with rigorous anonymization and no user re-identification. We further discuss data handling, consent considerations, and the broader ethical implications of persona-based modeling in the Appendix Sec. H.

References

1. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? (2022), <https://arxiv.org/abs/2105.14491>
2. Chen, Z., Zhang, L., Weng, F., Pan, L., Lan, Z.: Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7727–7736 (2024)
3. DeepMind, G.: Gemini 2.5 flash image (nano banana): Multimodal diffusion for image generation. <https://aistudio.google.com> (2025)
4. Deming, W.E., Stephan, F.F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* **11**(4), 427–444 (1940)
5. Du, B., Guo, M., He, S., Ye, Z., Zhu, X., Su, W., Zhu, S., Zhou, Y., Zhang, Y., Ai, Q., et al.: Twinvoice: A multi-dimensional benchmark towards digital twins via llm persona simulation. arXiv preprint arXiv:2510.25536 (2025)
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
7. Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., Yu, D.: Scaling synthetic data creation with 1,000,000,000 personas. arXiv preprint arXiv:2406.20094 (2024)
8. Henrich, J., Heine, S.J., Norenzayan, A.: The weirdest people in the world? *Behavioral and brain sciences* **33**(2-3), 61–83 (2010)
9. Inc., A.: Adobe firefly: Generative ai for creative content. <https://www.adobe.com/products/firefly> (2024)
10. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation (2024), <https://arxiv.org/abs/2401.09603>
11. Jung, S.G., Salminen, J., Aldous, K.K., Jansen, B.J.: Personacraft: Leveraging language models for data-driven persona development. *International Journal of Human-Computer Studies* **197**, 103445 (2025)
12. Kim, H., Ahn, S., Seo, Y.D.: Draw your mind: Personalized generation via condition-level modeling in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17171–17180 (2025)
13. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems* **36**, 36652–36663 (2023)
14. Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., Dong, X.: Better zero-shot reasoning with role-play prompting. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 4099–4113 (2024)
15. Lee, J., Heo, R., Lee, D.: Personalized reward modeling for text-to-image generation. arXiv preprint arXiv:2511.19458 (2025)
16. Ng, M.T., Tse, H.T., Huang, J.t., Li, J., Wang, W., Lyu, M.R.: How well can llms echo us? evaluating ai chatbots’ role-play ability with echo. arXiv preprint arXiv:2404.13957 (2024)
17. OpenAI: text-embedding-3-large. <https://platform.openai.com/docs/models/text-embedding-3-large> (2024), accessed: 2025-11-14

18. Park, S., Choi, S., Park, H., Yun, S.: Steering guidance for personalized text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15907–15916 (2025)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), <https://arxiv.org/abs/2103.00020>
20. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092 (2021)
21. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
22. Salehi, S., Shafiei, M., Yeo, T., Bachmann, R., Zamir, A.: Viper: Visual personalization of generative models via individual preference learning (2024), <https://arxiv.org/abs/2407.17365>
23. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? In: International conference on machine learning. pp. 29971–30004. PMLR (2023)
24. Shin, J., Hedderich, M.A., Rey, B.J., Lucero, A., Oulasvirta, A.: Understanding human-ai workflows for generating personas. In: Proceedings of the 2024 ACM Designing Interactive Systems Conference. pp. 757–781 (2024)
25. Singh, S.K., I, H.S., Singla, Y.K., Chen, C., Shah, R.R., Baths, V., Krishnamurthy, B.: Teaching human behavior improves content understanding abilities of VLMs. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=ff2V3UR9sC>
26. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
27. Tang, Y., Wang, B., Fang, M., Zhao, D., Huang, K., He, R., Hou, Y.: Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. arXiv preprint arXiv:2305.11482 (2023)
28. Von Rütte, D., Fedele, E., Thomm, J., Wolf, L.: Fabric: Personalizing diffusion models with iterative feedback. In: European Conference on Computer Vision. pp. 385–400. Springer (2024)
29. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
30. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis (2023), <https://arxiv.org/abs/2306.09341>
31. Xu, Y., Wang, W., Zhang, Y., Tang, B., Yan, P., Feng, F., He, X.: Personalized image generation with large multimodal models. In: Proceedings of the ACM on Web Conference 2025. pp. 264–274 (2025)
32. Yang, F., Chen, Z., Jiang, Z., Cho, E., Huang, X., Lu, Y.: Palr: Personalization aware llms for recommendation. arXiv preprint arXiv:2305.07622 (2023)
33. Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L., Leskovec, J.: Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 974–983. KDD '18, ACM (Jul 2018). <https://doi.org/10.1145/3219819.3219890>, <http://dx.doi.org/10.1145/3219819.3219890>

34. Zhou, Y., Zhu, Q., Jin, J., Dou, Z.: Cognitive personalized search integrating large language models with an efficient memory mechanism. In: Proceedings of the ACM Web Conference 2024. pp. 1464–1473 (2024)