Visual Personalization Turing Test

Rameen Abdal
Snap Research

James Burgess
Stanford University

Sergey Tulyakov
Snap Research

Kuan-Chieh Jackson Wang
Snap Research

https://snap-research.github.io/vptt

Abstract

We introduce the Visual Personalization Turing Test (VPTT), a new paradigm for evaluating contextual visual personalization based on perceptual indistinguishability, rather than identity replication. A model passes the VPTT if its output (image, video, 3D asset, etc.) is indistinguishable to a human or calibrated VLM judge from content a given person might plausibly create or share. To operationalize VPTT, we present the VPTT Framework, integrating a 10k-persona benchmark (VPTT-Bench), a visual retrieval-augmented generator (VPRAG), and the VPTT Score, a text-only metric calibrated against human and VLM judgments. We show high correlation across human, VLM, and VPTT evaluations, validating the VPTT Score as a reliable perceptual proxy. Experiments demonstrate that VPRAG achieves the best alignment–originality balance, offering a scalable and privacy-safe foundation for personalized generative AI.

1. Introduction

Personalization in visual generation has so far focused on *identity replication* [2–4, 6, 15, 16, 31, 34, 44, 45, 53], optimizing models to reproduce a subject across scenes. While effective at preserving appearance, these pipelines are computationally expensive [4, 15, 44] and miss the broader vision of personalization: *how individuals perceive, stylize, and share their world*. To instantiate this idea, personalization should capture the aesthetic preferences [39, 47, 54], cultural context, and visual familiarity that constitute a person's unique visual language. Yet, no benchmark exists to measure whether a model's output truly *feels like it could have been created by a particular person or a creator*. This gap is increasingly important beyond research. Industry is actively trying to bridge the gap between GenAI and user-created content to make generative AI *monetizable, trustworthy, and personally resonant* [11, 36]. This challenge becomes even more pressing as powerful foundation models in image domain, such as Qwen [58], NanoBanana [22] and GPT-Image-1 [37], already achieve near-photorealistic quality. As models master realism, the frontier of innovation shifts to what is personally relevant to the user [38].

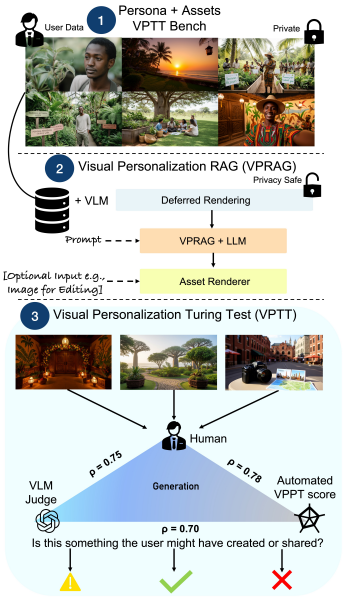

Figure 1. **Visual Personalization Turing Test.** We present the Visual Personalization Turing Test (VPTT), a new paradigm for contextual personalization at scale. A model passes the VPTT if its output is indistinguishable to a human or a calibrated VLM judge from what a given person might plausibly create or share. As one way to address this challenge, we introduce *VPTT Framework* consisting of privacy-safe benchmark *VPTT-Bench* for evaluating personalized generation and editing, and Visual Personalization RAG (*VPRAG*) that retrieves persona-aligned visual cues and converts them into personalized image generations or edits. To close the loop, we propose an automated $\mathrm{VPTT}_{\mathrm{score}}$ that achieves strong Spearman rank correlation (ρ) with humans and VLM Judges, establishing it as a cheap, reliable proxy for human perception of personalization.

To address this gap, we introduce the Visual Personalization Turing Test (VPTT) (Figure. 1): a new paradigm for evaluating generative models. A model passes the VPTT if its output (image, video, 3D asset etc.) is *indistinguishable to a human or a calibrated VLM judge from that a given*

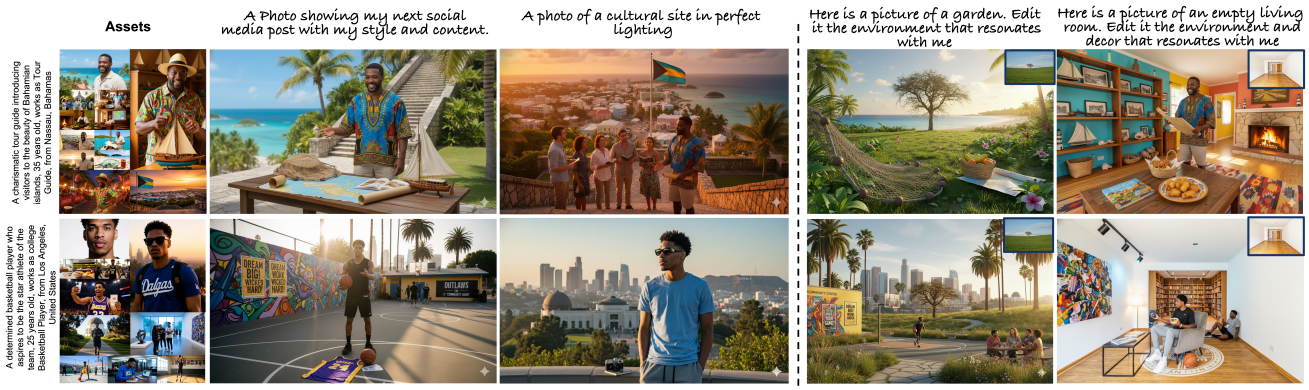

Figure 2. **Contextual Image Generation and Editing using VPTT-Bench.** Each row shows a distinct user profile: assets and style cues (left), personalized generations (social post, cultural site), and edits (garden, living room) guided by the same persona identity. All images are generated synthetically via our Visual Personalization RAG (VPRAG) by text, which retrieves persona-aligned cues. To show cross model personalization here the assets are generated by QWEN-image-model [58] and generations and edits by Nano-Banana [22] conditioned only on the first image. More results in are in Supplementary materials.

person might plausibly create or share. This reframes the goal from rote memorization of appearance to the far more challenging task of simulating a personal perspective.

Solving the VPTT presents three fundamental challenges. First, it requires a benchmark with thousands of diverse, culturally, and stylistically rich user profiles, yet real-world user data is inaccessible due to privacy concerns, fundamentally limiting academic research. Second, it demands a new technical approach beyond the fine-tuning one that can interpret a user's complex, multi-faceted style from their history and apply it to new generations in a scalable, efficient manner. Third, it requires a robust evaluation protocol to test VPTT at large scale.

We introduce the *VPTT framework*, designed to address these challenges at scale. To overcome the data barrier, we construct *VPTT-Bench*, the first large-scale benchmark of about 10,000 synthetic personas, whose visual worlds (30 assets - images for the scope of this paper) are represented entirely in text as "deferred renderings," (structured, attribute-rich intermediates like lighting , materials, environment, actions, forground, background, appearance etc. that defer visual realization, analogous to G-buffers [12] in graphics) enabling privacy-safe research at scale. Additionally, we render about 1000 synthetic personas to create a rich visual library. As a possible solution to personalization at scale, we propose a novel visual personalization retrieval-augmented generation (*VPRAG*) system. Instead of costly retraining, our method conditions generation on a persona's existing assets through hierarchical semantic retrieval with an optional learnable feedback and composes a personalized prompt enriched with their unique stylistic elements.

Our evaluation framework for image generation and editing is two fold. We first introduce $\mathrm{VPTT}_{\mathrm{score}}$ as a automatic proxy for VPTT. We conduct a visual-level evaluation through VPTT, validated by human study and extended with calibrated VLM judges. This helps us establish strong correlations among all three evaluators text-level ($\mathrm{VPTT}_{\mathrm{score}}$), VLM, and human, confirming that the $\mathrm{VPTT}_{\mathrm{score}}$ is a reliable, perceptually grounded proxy for visual judgment. After establishing this, we perform a large-scale deferred rendering analysis (about 120,000 evaluations) using the $\mathrm{VPTT}_{\mathrm{score}}$. Our results show that VPRAG's structured design achieves the best trade-off between output alignment and novelty, addressing a key limitation of black-box baselines. Our contributions are:

- A new task formulation, the Visual Personalization Turing Test (VPTT), redefines success in visual personalization as achieving human indistinguishable authenticity.
- VPTT Framework, the first scalable, privacy-safe benchmark for contextual personalization, featuring 10,000 rich personas with 1,000 visually rendered agents.
- A novel Visual Personalization Retrieval-Augmented Generation (VPRAG) system, a structured, zero-shot engine for personalization offering a possible scalable solution.
- A rigorous new evaluation framework featuring the VPTT score validated against human and VLM judges, proving it is a reliable proxy for perceptual alignment.
- A comprehensive analysis on our benchmark using a mix of closed- and open-source models with varying computational budgets, demonstrating that VPRAG offers a better trade-off between performance and efficiency.

2. Related Work

2.1. Personalization in Visual Generative Models.

Personalization in generative models has traditionally focused on identity replication [2–4, 6, 14, 15, 17, 31, 44]. Seminal methods like DreamBooth [44] and LoRA adaptations [46] excel at fine-tuning models to reproduce a specific subject across different scenes. However, these approaches are not scalable and primarily address appear-

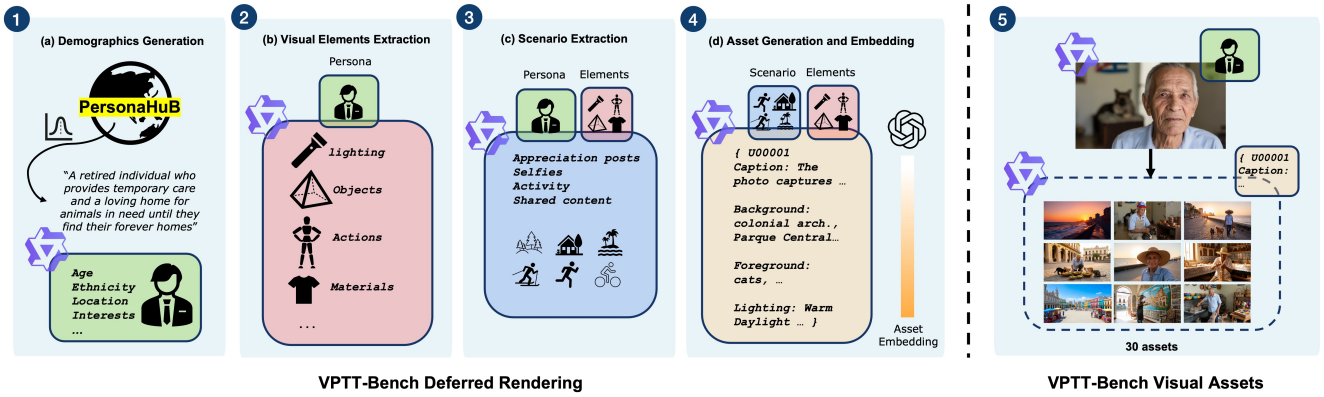

Figure 3. **VPTT-Bench Data Generation Pipeline.** Overview of the deferred rendering pipeline used to construct VPTT-Bench. (1) Personas are sampled from PersonaHub [21] with demographics. (2–3) Visual and scenario elements (lighting, actions, materials etc.) are extracted. (4) These cues are composed into structured captions and embedded via an LLM. (5) Generating 30 corresponding visual assets per persona, forming privacy-safe, semantically grounded data for evaluating contextual personalization.

Figure 4. **Example Personas from VPTT-Bench.** Each row shows a synthetic persona sampled from PersonaHub [21] (only short descriptions) with its corresponding visual assets generated via VPTT-Bench generation pipeline. Personas span diverse regions, professions, and age groups, illustrating the demographic and contextual diversity of VPTT-Bench.

ance fidelity rather than the user's broader visual signature [2, 14–17, 31, 34, 43–45, 49, 52, 53, 56]. More recent works aim for tuning-free personalization. IP-Adapter and related works in the image and video domains [3, 6, 17, 43, 49, 56, 59] use reference images to condition generation, achieving strong results in transferring style or appearance [13, 19, 25, 26] but often requiring careful selection of reference images and suffer from the absence of a larger visual context [19]. Methods like InstantBooth [50] represent another direction in test-time personalization without fine-tuning but again focuses on personalizing the appearance of the subject. Among the methods that consider the context, DrUM [29] proposes learning a vector based on prompt history and injecting it via a trained adapter network, offering a modular approach but still involving per user adapter training. A very recent work ImageGem [23], collects in-the-wild interactions for generative model personalization, highlighting the community's growing interest in this area, though primarily focused on LoRAs collected over users generated content. Our work, orthogonal to these works, focuses on deriving and applying preferences, cultural context, visual familiarity and personal elements implicitly derived from a user's asset history, without requiring explicit reference images or per-user training of adapters.

2.2. Visual Preference Personalization

Aligning generative models with user preferences is a critical challenge. Many recent efforts draw inspiration from Reinforcement Learning from Human Feedback (RLHF) [41] used in LLMs [9, 58]. An early work, ImageReward [57] trained a reward model on human comparisons to score prompt-image alignment, enabling fine-tuning via Reward Feedback Learning (ReFL). Diffusion-DPO [51] applied Direct Preference Optimization to fine-tune Stable Diffusion XL [42] on large-scale human judgments [55] from datasets like Pick-a-Pic [30], improving general appeal and alignment. While powerful, these methods typically optimize for aggregate preferences rather than individual context. On the other hand, approaches targeting individual preferences are emerging [35]. ViPer [47] learns preferences by having an MLLM [39] analyze user comments on images, extracting structured attributes to guide generation. PPD [10] trains a single model conditioned on user embeddings derived from few-shot pairwise preferences. POET [24] focuses on identifying image homogeneity using "prompt inversion" and personalizing diversification based on interactive user feedback. Concurrent work, such as Instant Preference Alignment [32], also uses

MLLMs to extract preferences from a reference image for tuning-free guidance. Our work differs by focusing on extracting and applying alignment implicitly from a user's historical creative output (simulated via *VPTT-Bench* derived from real-world grounded PersonaHuB [21]) rather than relying on explicit feedback, pairwise comparisons, or single reference images. We introduce the VPTT as a holistic measure of visual context alignment beyond simple preference scores.

2.3. RAG in Computer Vision

Retrieval-Augmented Generation (RAG) [20], initially prominent in NLP, is increasingly being explored in computer vision [48, 61]. Very recent works like RealRAG [33] and FineRAG [60] focused on retrieving external visual knowledge (e.g., real images of objects) to improve content completion of generated images and using RAG for VQA. Comprehensive repositories like Awesome-RAG-Vision [61] are mapping the growing landscape, covering applications in visual understanding, generation, and embodied AI. Within generation, RAPO [18] uses RAG specifically for text-to-video prompt optimization, retrieving terms from a graph built on training data to align user prompts with the model's expected input format. Tailored Visions [7] pioneered using RAG on a user's own prompt history for personalized text-to-image prompt rewriting, using an LLM to synthesize past styles into new prompts. OmniStyle [54], while focused on style transfer, utilizes a large curated dataset and filtering for high-quality supervised training. Our VRAG system builds upon the personalized RAG concept but distinguishes itself through: (1) operating on our structured, synthetic *VPTT-Bench* benchmark, enabling privacy-safe research; and (2) employing a principled, more transparent retrieval and composition architecture for fine-grained control, rather than relying solely on a black-box LLM operating on raw prompt history.

3. Visual Personalization Turing Test

Our goal is to model and evaluate *contextual visual personalization* the ability of a generative model to produce content that a human (or VLM) would perceive as consistent with a given persona's visual context. We formalize this as the Visual Personalization Turing Test (VPTT) and introduce VPTT Framework, a unified framework that enables systematic study of this problem at scale. VPTT Framework consists of four interacting components: (1) a large-scale simulated persona benchmark (Sec. 3.1); (2) a retrieval-augmented generation engine (Sec. 3.2); (3) an optional learnable feedback loop (Sec. 3.3); and (4) a differentiable proxy metric, VPTT score (Sec. 3.4). Together they form a closed cycle of simulation → generation → judgment → optimization.

Problem Definition. Given a persona $\mathcal{P} = \{d, E, C\}$ demographics d, a structured element library E, and a caption memory C and a query p, the model must generate a personalized prompt p' whose resulting image $\mathcal{G}(p')$ maximizes perceived alignment with $\mathcal{P}$:

$$\mathcal{J}(p'; \mathcal{P}) = \lambda_1 \,\text{Align}(p', \mathcal{P}) + \lambda_2 \,\text{Fidelity}(p', C) + \lambda_3 \,\text{Novelty}(p', C), \qquad \sum_i \lambda_i = 1. \tag{1}$$

This surrogate defines the latent VPTT objective: an ideal system achieves high alignment, high fidelity, and high novelty simultaneously, an intractable trade-off for current models. We expect this trade-off to improve with better personalized models and for the scope of this work propose a method that approximates this objective efficiently without retraining.

3.1. VPTT-Bench: Scalable Simulation Substrate

Human personalization datasets are private and unscalable. We therefore construct **VPTT-Bench** (Figure. 3 and Figure. 4), a synthetic benchmark of 10,000 agents, each represented by a tuple $\mathcal{P}_i = \{d_i, E_i, C_i\}$. Personas are generated using `Qwen2.5-72B-Instruct` [58]:

Demographic Generation: starting from public textual seeds (PersonaHUB [21]), we sample culturally diverse backstories d_i. This ensures cross-domain coverage, avoiding dataset bias.

Visual Elements Extraction: we sample and cluster atomic visual terms (e.g., clothing, lighting, pose) into structured vocabularies E_i conditioned on d_i ensuring the visual elements are consistent with the persona.

Scenario and Assets Extraction: conditioned on $\{d_i, E_i\}$, we first generate short scenarios of the assets and finally generate 30 captions C_i describing element rich posts with the scenario story arc. The captions are embedded using `text-embedding-3-small` [39].

We further render a 1,000-persona subset into image galleries (30 images per persona), each anchored by a canonical portrait followed by caption-guided edits. This hybrid text–image corpus provides both semantic control and visual diversity: the text-only component enables dense, scalable supervision without privacy constraints, while the paired visual assets allow controlled studies across different resource budgets, from lightweight text-only personalization to more expensive multimodal (text + image) setups. For real profiles, the reverse of this process is performed to get the structured data.

3.2. Visual Personalization Retrieval-Augmented Generation (VPRAG)

To personalize content without model retraining, we propose **VPRAG** (see Figure. 5), a retrieval-augmented generation framework that conditions prompt rewriting on a

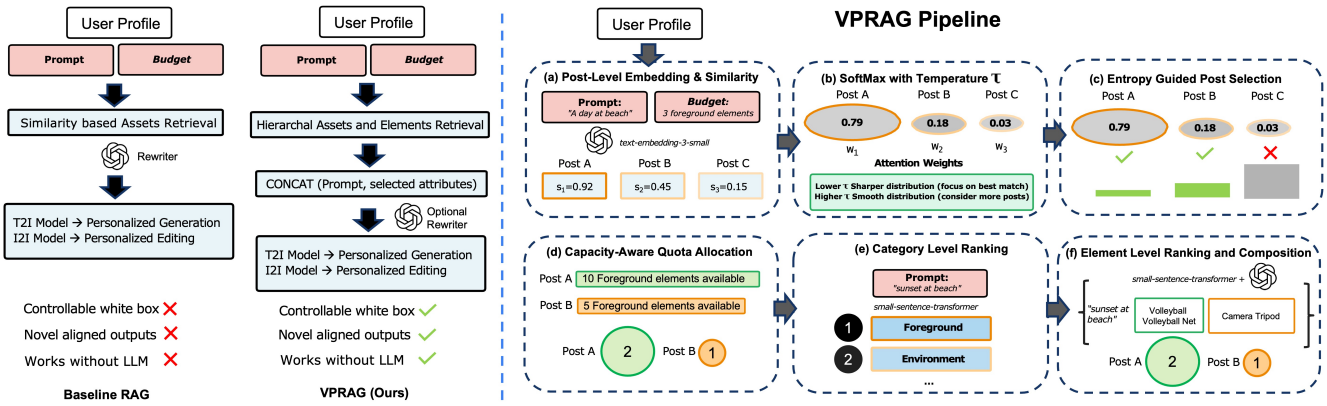

Figure 5. **VPRAG Pipeline Overview.** Comparison between the baseline retrieval-augmented generation (BRAG) and our proposed Visual Personalization RAG (VPRAG). Unlike baseline BRAG, VPRAG introduces controllable and interpretable retrieval through: (a) post-level embedding and similarity scoring, (b) temperature-controlled attention, (c) entropy-guided post selection, (d) capacity-aware quota allocation, (e) category-level ranking, and (f) element-level composition. This multi-stage design yields a white-box, LLM-optional retrieval framework producing visually and semantically aligned personalized generations and edits.

persona's structured memory. Given a query p and profile $\mathcal{P} = \{d, E, C\}$, VPRAG retrieves semantically relevant posts and elements, allocates retrieval quotas, and composes a new prompt p' that aligns with the persona's context. Unlike other methods [1, 44] that require minutes to hours per user, VPRAG operates entirely at inference time, adding only a few hundred milliseconds of retrieval and composition overhead.

Hierarchical Retrieval. Captions C encode holistic semantic intent (high-level concepts), while elements E capture atomic style (low-level cues). We therefore perform a hierarchical two-level retrieval for robustness.

Post-level retrieval. Each persona's captions $\{c_i\}$ are embedded using `text-embedding-3-small` [39], and cosine similarities $s_i = \mathbf{q}^\top \mathbf{v}_i$ are computed with the query p. Weights are normalized as $w_i = \frac{\exp(s_i/\tau)}{\sum_j \exp(s_j/\tau)}$, where τ is a softmax temperature controlling retrieval sharpness. This Boltzmann weighting represents the *maximum-entropy solution* for expected semantic alignment under a temperature constraint [28], guaranteeing smooth attention while avoiding brittle hard cutoffs.

Entropy Guided post Selection. We then measure entropy $H = -\sum_i w_i \log w_i, \quad n_{\text{eff}} = \exp(H)$, where n_{eff} approximates the *effective number of relevant posts*,a theoretically grounded proxy for query specificity. Broader prompts (e.g., "in the park") yield higher H and therefore encourage more diverse retrieval, whereas narrower ones (e.g., "in Kashmiri traditional dress") produce lower entropy, focusing the selection. To balance adaptivity and efficiency, we cap the retrieved posts given the budget Q (total number of visual elements to sample from categories $\mathcal{C} = \{\text{fg}, \text{bg}, \text{lighting}, \text{pose}, \ldots\}$), set as $K = \min(\lfloor n_{\text{eff}} \rfloor, 2 \times Q)$, ensuring controlled expansion without over-retrieval for broad prompts.

Quota Allocation. Each post contributes elements from categories $\mathcal{C}$. Given category $c' \in \mathcal{C}$, we allocate quotas to each post i as: $q_i^{(c')} = \left\lfloor \frac{w_i \cdot n_i^{(c')}}{\sum_j w_j \cdot n_j^{(c')}} \cdot Q_{c'} \right\rfloor$ where $n_i^{(c')}$ is the number of available elements in category c' for post i, and $Q_{c'}$ is the total budget for category c'. Remainders are allocated to largest–fraction posts. This rule ensures the proportional-fair allocation objective so that high-weight posts get more samples, but low-weight ones still contribute diversity.

Element-level retrieval. Within the top-K posts we prioritize the categories based on the prompt p using semantic relevance $\text{score}_k = \cos(\phi(\mathbf{c}_k), \phi(p))$, ($\phi$ is a lightweight transformer encoder (`MiniLM`) [27]). Within each category, elements are ranked based on the closeness to the p using the same `MiniLM` [27], and the top-$q_i^{(k)}$ are selected.

Prompt Composition. The selected elements $\mathcal{E}_p$ are concatenated with persona summary $\mathcal{S}_p$ into $p' = f_{\text{compose}}(p, \mathcal{S}_p, \mathcal{E}_p, L)$ under a token-length budget L. This yields a re-prompt enriched with stylistic and contextual cues consistent with the persona's memory. Based on the budget, f_{compose}, can be an LLM refining the story arc for the generation or a simple text concatenation.

3.3. Learnable Feedback Simulation

While VPRAG uses persona aligned retrieval, personalization also involves subjective preference learning. We therefore introduce a small learnable feedback module to approximate user-specific value functions. Given persona $\mathcal{P}$ with subjective preferences and generated prompt p', a vision–language judge (VLM) outputs an alignment score $s_{\text{VLM}} \in [0, 1]$. We train a cross-attention predictor f_θ to estimate $\hat{s}_{\text{VLM}} = f_\theta(\text{Emb}(p'), \text{Emb}(\mathcal{P}))$, and re-rank candidates by $p'^* = \arg\max_m f_\theta(\text{Emb}(p'_m), \text{Emb}(\mathcal{P}))$. We

use this component as a smaller scale proof of concept to encourage future extensions of VPTT Framework toward closed-loop personalization.

3.4. VPTT Score: A Differentiable Proxy for Personalization

We now introduce $\mathrm{VPTT}_{\mathrm{score}}$, a quantitative metric that serves as the text-level scalable foundation for the VPTT triangle and a convex surrogate of the personalization objective in Eq. 1. $\mathrm{VPTT}_{\mathrm{score}}$ combines four interpretable metrics that jointly approximate alignment, fidelity, and originality: *Persona Alignment (PA)*, *GS Reconstruction (GS)*, *Cluster Proximity (CP)*, and *Novelty (NV)*.

(1) Persona Alignment (PA). This term measures semantic coherence between the generated prompt p' and the textual description of the persona $\mathcal{P}$: $\mathrm{PA}(p', \mathcal{P}) = \cos\big(\mathrm{Emb}(p'), \mathrm{Emb}(\mathcal{P})\big)$.

(2) GS Reconstruction (GS). To measure content fidelity, we represent each persona's caption embeddings $\{v_i\}$ as an orthonormal basis B using the Gram–Schmidt process. For a generated prompt embedding v_p, $\mathrm{GS}(p', C) = \cos\big(v_p,\, B(B^\top v_p)\big)$ which evaluates how well p' can be reconstructed from the assets's semantic span. GS measures subspace fidelity i.e. whether a generation stays within the semantic manifold defined by the persona's gallery rather than mere pairwise similarity.

(3) Cluster Proximity (CP). To assess thematic consistency, all asset captions are clustered in the GS basis thematic centroids $\{c_k\}$. The hard version used for evaluation is $\mathrm{CP}(p', C) = \exp\big(-\min_k \|v'_p - c_k\|_2\big)$, while the differentiable relaxation replaces min with a temperature-controlled softmin: $\widetilde{\mathrm{CP}}(p', C) = \sum_k \frac{\exp(-\|v'_p - c_k\|_2/\tau)}{\sum_j \exp(-\|v'_p - c_j\|_2/\tau)}$.

(4) Novelty (NV). Novelty penalizes verbatim reuse of retrieved captions. The discrete version measures maximum trigram overlap: $\mathrm{NV}(p', C) = 1 - \max_i \frac{|\mathrm{Tri}(p') \cap \mathrm{Tri}(c_i)|}{|\mathrm{Tri}(p')|}$. For differentiable analysis, we define a soft-overlap relaxation: $\widetilde{\mathrm{NV}}(p', C) = 1 - \max_i \frac{\sum_t \cos(\phi_t(p'), \phi_t(c_i))}{|\mathrm{Tri}(p')|}$, where $\phi_t(\cdot)$ denotes continuous n-gram embeddings (via small sentence transformer for example `MiniLM` [27]).

Combined Score. The overall proxy is a convex weighted combination: $\mathrm{VPTT}_{\mathrm{score}} = 0.20\,\mathrm{PA} + 0.30\,\mathrm{GS} + 0.30\,\mathrm{CP} + 0.20\,\mathrm{NV}$. Empirically, GS and CP correlate most strongly with human visual fidelity, so we assign them higher weight (0.3 each). PA measures semantic alignment (0.2), while NV promotes originality and prevents overfitting (0.2). The weighting satisfies $\sum_i \lambda_i = 1$, forming an unbiased convex estimator of $\mathcal{J}$. For tasks with limited prompt budgets (e.g., adding exactly three retrieved phrases), the novelty term becomes less meaningful as textual overlap is bounded by design. We therefore use the normalized variant $\mathrm{VPTT}_{\mathrm{score}}\text{-c} = \frac{1}{3}(\mathrm{PA} + \mathrm{GS} + \mathrm{CP})$, which equally weighs the three active components. We further justify the weights in Sec 4.2.1 while computing the correlations. The novelty term is also set to zero for the baselines not conditioned on the captions. While our experiments report the hard (evaluation) forms for interpretability, the differentiable variant makes $\mathrm{VPTT}_{\mathrm{score}}$ suitable as a learnable objective in future personalization pipelines.

4. Evaluations

4.1. Baselines

We benchmark VPTT Framework against two baseline categories. First, scalable privacy-safe pipelines including *Baseline* - no access to any asset, *Persona Only* - access to demographics information, and Baseline RAG *BRAG* [7], a strong baseline with access to all the persona captions for personalization (see Figure. 5). These operate via retrieval and rewriting without model retraining, allowing large-scale evaluation across 10,000 personas. Second, we reference high-cost personalization baselines such as DB-LoRA [1], Flux [5], DrUM [29], MLLM [39, 58], and VIPER [47], which rely on user-specific fine-tuning or only preference optimization. These are computationally intensive and non-scalable, so we evaluate them only on smaller subsets and report results in the Supplementary. This separation highlights VPTT's focus on scalable, privacy-safe personalization while remaining comparable to existing high-fidelity methods.

4.2. Quantitative Evaluation

Evaluating the VPTT is intrinsically challenging because the outcome depends on a cascade of interacting systems:

1) **Prompt Generation:** The rewriter LLM must faithfully express a persona's stylistic intent.
2) **Image Generation:** The T2I or I2I model must accurately translate those prompts into coherent visual content.
3) **Evaluation:** The VLM judge must perceive the subtle consistency between the generated content and the persona's authentic visual identity.

VPTT performance improves as these three domains mature. To systematically evaluate them, we design a three-stage protocol addressing three central questions (**Q1–Q3**). All experiments are conducted across a spectrum of models from open-source **Qwen2.5-7B-Instruct** [58] to efficient **GPT-4o-mini** [40] and high-capacity **Gemini-2.5-Pro** [9] ensuring robustness across compute budgets. To make the evaluation holistic, we consider both image generation and editing tasks.

4.2.1. Q1: Can We Trust Our Metrics?

Before scaling the evaluation, we verify that our automated metrics i.e. VLM judgment and the text-only $\mathrm{VPTT}_{\mathrm{score}}$ faithfully approximate human perception.

Table 1. Quantitative comparison for generation and Editing Tasks across 6000 human annotations. We report mean (**Avg.**) and accuracy (**Acc.**) scores for three evaluation levels: text-based $\mathrm{VPTT}_{\mathrm{score}}$-c **(0–1)**, vision-language VLM (0–5), and human judgments Human (0–5). Higher is better for all.

Method	$\mathrm{VPTT}_{\mathrm{score}}$-c (Text)		VLM (Visual)		Human (Perceptual)	
	Avg.	**Acc.**	**Avg.**	**Acc.**	**Avg.**	**Acc.**
Baseline	0.329	0.0%	2.41	4.6%	1.64	0.70%
Persona Only	0.400	7.3%	3.32	19.2%	2.51	16.0%
BRAG	0.420	19.3%	3.52	21.6%	2.69	21.3%
VPRAG (Ours)	**0.464**	**73.3%**	**4.32**	**54.6%**	**3.34**	**62.0%**

Human Study. We collected about 6,000 human ratings using images across *four* methods (see Table. 1), *three* LLM generations and *two* tasks (image generation "A preferred outdoor spot" and editing "Here is a convention center. Add a preferred event"), from 20 annotators. Inter-annotator agreement was substantial (Kendall's $W = 0.651 \pm 0.141$ for Generation, 0.564 ± 0.209 for Editing), confirming consistent human understanding of personal authenticity.

Metric Calibration and Validity. We validate the proposed metrics by measuring Spearman's rank correlation (ρ) between automated judgments and human ratings (Figure 1). For efficient evaluation, we use 10 visually and semantically matched posts out of 30 under a budgeted evaluation setup (Sec. 3.4). We calibrate the VLM judges using GPT-4o and Gemini-2.5-Pro, wherever applicable to remove evaluation bias on a small set. In evaluation of the whole set, VLM-based judgments strongly align with human perception (combined $\rho = 0.67$, generation: 0.75). Our text-only $\mathrm{VPTT}_{\mathrm{score}}$-c metric achieves comparable agreement (combined $\rho = 0.68$, generation: 0.78) with a Top-2 agreement accuracy of 99%, confirming its reliability as a human-perceptual proxy. $\mathrm{VPTT}_{\mathrm{score}}$-c also correlates well with VLM scores (combined $\rho = 0.57$, generation: 0.70), indicating consistent cross-modal alignment. While editing correlations are lower ($\rho \approx 0.5$) due to the finer granularity of localized visual edits and potential perceptual losses after downsampling, generation consistently exceeds 0.7, demonstrating the robustness of our metric design. Finally, we report the averaged raw scores in Table. 1 where our method VPRAG is a clear winner across all the evaluations. Overall, these results establish $\mathrm{VPTT}_{\mathrm{score}}$-c as a fast, low-cost, and perceptually grounded surrogate for human evaluation in large-scale personalization studies.

4.2.2. Q2: Does a Better Prompt Create a Better Image?

With calibrated evaluators, we conduct the main VPTT experiment on 200 personas on *two* tasks (across *three* LLM models and *five* methods) under a fixed "three-phrase budget" to ensure fair comparison. This part disentangles what visual generation is able to achieve with models' ability to generate authentic detailed prompts (we evaluate that next). Evaluation of this extended dataset mirrors the correlation $\rho = 0.53$ (generation : 0.66) of the previous section. Table. 2 shows the results (averaged across LLMs) for the generation and editing tasks. The evaluation again shows how hierarchical controllable retrieval does not confuse the models and produce better alignments.

Table 2. Comparison of Generation and Editing tasks on 200 personas after VLM calibration across 3 LLM rewrite methods. We report mean $\mathrm{VPTT}_{\mathrm{score}}$-c (V-c) and VLM scores along with winning accuracy (%). Higher is better.

Method	Generation				Editing			
	V-c	Acc.	VLM	Acc.	V-c	Acc.	VLM	Acc.
Baseline	0.343	0.0%	2.21	1.4%	0.322	0.0%	2.97	10.5%
Persona Only	0.402	1.2%	2.98	5.9%	0.399	9.2%	3.44	18.5%
BRAG	0.451	18.4%	4.04	25.6%	0.415	15.3%	3.75	24.3%
VPRAG (Ours)	**0.472**	**41.7%**	4.08	31.0%	**0.448**	**47.2%**	**4.03**	**30.8%**
Comb. (Ours)	0.472	38.8%	**4.30**	**36.1%**	0.436	28.3%	4.03	15.8%

4.2.3. Q3: Is the Architecture Robust at Scale?

To assess generalization, we evaluate all models text-only across our entire VPTT-Bench benchmark of 10,000 personas and four tasks (*two* generation, *two* editing , see Figure. 2), totaling 120,000 prompt evaluations. The prompts are limited to 150 words and a budget of 3 ($\tau = 0.1$) is allocated to all visual element Categories $\mathcal{C}$. The elements are arranged in decreasing order of relevance and LLM is given freedom to choose from the list to orchestrate a story arc. As shown in Table 3, naive rewriters (BRAG) overfit to captions (often copy-pasting them), earning high alignment but low originality scores (more detailed in Supplementary) and hence falling short. In contrast, VPRAG consistently achieves the best composite $\mathrm{VPTT}_{\mathrm{score}}$, maintaining the optimal balance between alignment and originality across all rewriter backbones. This large-scale experiment demonstrates that VPRAG scales linearly, generalizes across models, and sustains perceptual authenticity without retraining.

4.2.4. Downstream Study: Feedback Simulation

We evaluate feedback simulation on a smaller subset of 200 personas (10,000 labeled examples) as a proof of concept rather than a core benchmark. Although this component is not used in our main quantitative evaluations, it demonstrates that compact models can learn to simulate user-level preference alignment from limited supervision. We sample diverse simulated profiles (95% occupation uniqueness, 96 countries, 10 ethnicity groups) and use GPT-4o [40] to generate 50 labeled prompts per profile, 20 aligned, 20 misaligned, and 10 neutral, yielding 10,000 labeled examples with profile-level splits (130/20/50 train/val/test). A compact cross-attention regressor (128-dim, 4 heads) achieves 73.8% overall accuracy (MAE: 0.1259) and 91.6% accuracy on aligned preference predictions for 50 unseen users (2,525 prompts), with only a 0.7% validation–test gap, showing that compact models can effectively capture persona-aware preferences while generalizing to new users. We leave large

Table 3. Main text-level results across 10,000 personas and three LLM models. We report the novelty-adjusted $\mathrm{VPTT}_{\mathrm{score}}$ (**V**), plus Cohen's **d** [8] (**d** $= \frac{|\mu_{\mathrm{best}} - \mu_{\mathrm{method}}|}{s_{\mathrm{pooled}}}$), measuring effect size relative to the best-performing method per row (μ_{best}) across 20,000 samples per entry. **Bold** indicates the best method and underline the second-best. The *Baseline* and *Persona Only* methods consistently underperform across both generation and editing tasks. Our *VPRAG* and *Comb.* (BRAG + VPRAG) methods achieve the best overall performance, with *Comb.* performing slightly better for 4o-mini (GPT-4o-mini [39]) and Gemini (Gemini-2.5-pro [9]), while *VPRAG* excels for Qwen (Qwen2.5-7B-Instruct [58]). Higher Cohen's **d** values ($d \geq 0.5$ indicates medium to large effects) demonstrate substantial performance differences, particularly between persona-based methods and baselines. See supplementary material for detailed score breakdowns.

(a) Generation

	Baseline		Persona Only		BRAG		VPRAG		Comb.	
Model	**V**	***d***	**V**	***d***	**V**	***d***	**V**	***d***	**V**	***d***
Qwen	0.316	11.9	0.389	8.3	0.581	1.1	**0.631**	NA	0.602	0.7
4o-mini	0.316	12.6	0.402	8.4	0.628	0.5	0.640	0.1	**0.644**	NA
Gemini	0.316	9.8	0.379	7.1	0.616	0.3	0.625	0.2	**0.632**	NA

(b) Editing

	Baseline		Persona Only		BRAG		VPRAG		Comb.	
Model	**V**	***d***	**V**	***d***	**V**	***d***	**V**	***d***	**V**	***d***
Qwen	0.306	12.0	0.378	8.7	0.583	1.1	**0.626**	NA	0.586	1.0
4o-mini	0.306	12.0	0.384	8.8	0.596	0.9	**0.626**	NA	0.610	0.5
Gemini	0.306	10.7	0.372	8.1	0.583	0.6	0.605	0.0	**0.606**	NA

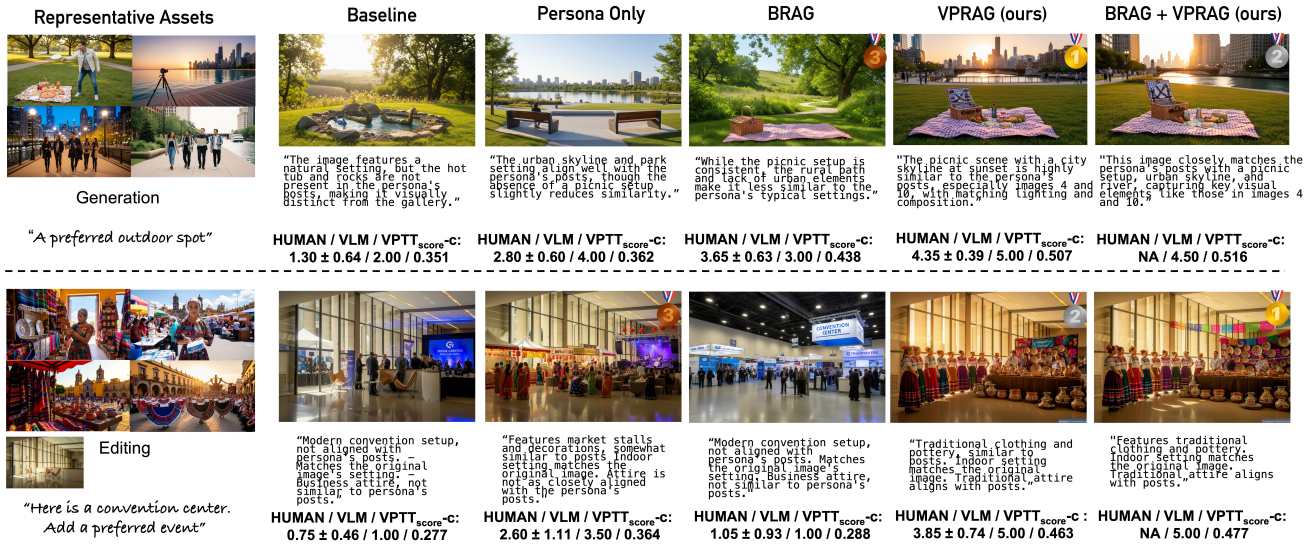

Figure 6. **Qualitative Comparison across Generation and Editing Tasks.** Representative examples from the VPTT-Bench showing outputs from five methods: Baseline, Persona Only, BRAG, VPRAG (ours), and BRAG + VPRAG (ours). Each sample is evaluated using human, VLM (reasoning shown), and text-level $\mathrm{VPTT}_{\mathrm{score}}$-c scores, where higher indicates closer alignment to the persona's assets. Our methods achieve the highest perceptual and text–visual consistency, confirming effective contextual personalization.

scale studies to future extensions.

4.3. Qualitative Results

VPRAG produces visually coherent and persona-faithful generations across diverse profiles. By retrieving fine-grained visual cues such as lighting, attire, scene semantics, and stylistic markers, VPRAG enriches the composed prompts while preserving originality and user-specific visual elements (Figure 2). These examples also highlight VPRAG's ability to perform cross model personalization, where VPRAG produces consistent personalization across QWEN-Image [58] and Nano-Banana [22].

Compared to the persona-only baseline (Figure 6) and the BRAG baseline, VPRAG achieves stronger contextual grounding, sharper visual fidelity, and more consistent preservation of persona style. For editing tasks, it additionally injects semantically relevant visual elements. Results for remaining baselines and additional profiles are provided in the Supplementary.

5. Conclusion

We introduced the Visual Personalization Turing Test (VPTT) as a principled paradigm for evaluating contextual visual personalization, and proposed the VPTT Framework, a scalable system that operationalizes this paradigm. The framework integrates VPTT-Bench, the VPRAG retrieval engine, and the $\mathrm{VPTT}_{\mathrm{score}}$ metric into a closed-loop pipeline for simulation, generation, and evaluation without any per-user retraining. Our results show strong alignment among human judgments, VLM judges, and the text-only $\mathrm{VPTT}_{\mathrm{score}}$, validating the framework as an efficient, privacy-safe foundation for personalized generative models. Future work will incorporate opt-in and federated real-user signals to further bridge simulated and real personalization while preserving user privacy.

References

[1] Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2022. 5, 6

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 1, 2, 3

[3] Rameen Abdal, Or Patashnik, Ekaterina Deyneka, Hao Chen, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Zero-shot dynamic concept personalization with grid-based lora, 2025. 3

[4] Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 1, 2

[5] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 6

[6] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Ivan Skorokhodov, Jun-Yan Zhu, Kfir Aberman, Ming-Hsuan Yang, and Sergey Tulyakov. Videoalchemy: Open-set personalization in video generation, 2024. 1, 2, 3

[7] Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting, 2024. 4, 6

[8] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013. 8

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 3, 6, 8

[10] Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models, 2025. 3

[11] Google DeepMind. Veo2. *https://deepmind.google/technologies/veo/veo-2/*, 2024. 1

[12] Michael Deering, Stephanie Winner, Bic Schediwy, Chris Duffy, and Neil Hunt. The triangle processor and normal vector shader: a vlsi system for high performance graphics. *SIGGRAPH Comput. Graph.*, 22(4):21–30, 1988. 2

[13] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024. 3

[14] 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. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3

[15] 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. In *ICLR*, 2023. 1, 2

[16] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. In *Siggraph*, 2023. 1

[17] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization. *arXiv preprint arXiv:2404.03620*, 2024. 2, 3

[18] Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, and Yaohui Wang. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation, 2025. 4

[19] Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[20] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. 4

[21] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2025. 3, 4

[22] Google. Nanobanan. *https://aistudio.google.com/models/gemini-2-5-flash-image*, 2025. 1, 2, 8

[23] Yuanhe Guo, Linxi Xie, Zhuoran Chen, Kangrui Yu, Ryan Po, Guandao Yang, Gordon Wetztein, and Hongyi Wen. Imagegem: In-the-wild generative image interaction dataset for generative model personalization, 2025. 3

[24] Evans Xu Han, Alice Qian Zhang, Haiyi Zhu, Hong Shen, Paul Pu Liang, and Jane Hsieh. Poet: Supporting prompting creativity and personalization with automated expansion of text-to-image generation, 2025. 3

[25] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. 2023. 3

[26] Hexiang Hu, Kelvin C. K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, Ming-Wei Chang, and Xuhui Jia. Instruct-imagen: Image generation with multi-modal instruction, 2024. 3

[27] huggingface. Minilm-l6-v2. *https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2*, 2025. 5, 6

[28] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957. 5

[29] Hyungjin Kim, Seokho Ahn, and Young-Duk Seo. Draw your mind: Personalized generation via condition-level modeling in text-to-image diffusion models, 2025. 3, 6

[30] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 3

[31] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 1, 2, 3

[32] Yang Li, Songlin Yang, Xiaoxuan Han, Wei Wang, Jing Dong, Yueming Lyu, and Ziyu Xue. Instant preference alignment for text-to-image diffusion models, 2025. 3

[33] Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. Realrag: Retrieval-augmented realistic image generation via self-reflective contrastive learning, 2025. 4

[34] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 1, 3

[35] Ofir Nabati, Guy Tennenholtz, ChihWei Hsu, Moonkyung Ryu, Deepak Ramachandran, Yinlam Chow, Xiang Li, and Craig Boutilier. Preference adaptive and sequential text-to-image generation, 2025. 3

[36] OPENAI. Sora. *https://openai.com/sora/*, 2024. 1

[37] OPENAI. Gpt-image-1. *https://openai.com/*, 2025. 1

[38] OPENAI. Sora2. *https://openai.com/index/sora-2/*, 2025. 1

[39] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report, 2024. 1, 3, 4, 5, 6, 8

[40] OpenAI et al. Gpt-4o system card, 2024. 6, 7

[41] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 3

[42] 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, 2023. 3

[43] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 3

[44] 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 *CVPR*, pages 22500–22510, 2023. 1, 2, 5

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 1, 3

[46] Simo Ryu. Dreamboothlora, 2023. 2

[47] Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. ViPer: Visual personalization of generative models via individual preference learning. *arXiv preprint arXiv:2407.17365*, 2024. 1, 3, 6

[48] Rotem Shalev-Arkushin, Rinon Gal, Amit H. Bermano, and Ohad Fried. Imagerag: Dynamic image retrieval for reference-guided image generation, 2025. 4

[49] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 3

[50] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning, 2023. 3

[51] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. 3

[52] Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv preprint arXiv:2404.11565*, 2024. 3

[53] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. High-fidelity person-centric subject-to-image synthesis. *arXiv preprint arXiv:2311.10329*, 2023. 1, 3

[54] Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale, 2025. 1, 4

[55] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 3

[56] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 3

[57] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. 3

[58] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. 1, 2, 3, 4, 6, 8

[59] 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. 3

[60] Huaying Yuan, Ziliang Zhao, Shuting Wang, Shitao Xiao, Minheng Ni, Zheng Liu, and Zhicheng Dou. FineRAG: Fine-grained retrieval-augmented text-to-image generation. In

Proceedings of the 31st International Conference on Computational Linguistics, pages 11196–11205, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. 4

[61] zhengxuJosh. Awesome-rag-vision. *https://github.com/zhengxuJosh/Awesome-RAG-Vision*, 2025. 4