PREFGEO: PREFERENCE GEOMETRY FOR MULTI-PREFERENCE FUSION OF FLOW MATCHING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human preference alignment has emerged as an effective approach for improving flow-matching-based text-to-image generation. However, existing methods primarily focus on aligning models with a single preference, while extending alignment to multiple preferences, such as compositional fidelity, text rendering, and aesthetics, remains challenging. Recent training-based multi-preference alignment methods, including DiffusionOPD, DiffusionNFT, and MapReduce LoRA, require at least **80GB** of GPU memory and **170 GPU hours** of training, making personalized multi-preference alignment computationally expensive and difficult to scale. In this paper, we introduce **PrefGeo**, a training-free framework that exploits the *preference geometry* between preference-aligned models and their shared base model. PrefGeo fuses models aligned with different preferences into a unified model without additional training. The fusion involves only model-weight operations and can be completed on a single NVIDIA RTX 4090 24GB GPU in **18 minutes**, using only **0.5GB** of GPU memory. Extensive experiments demonstrate that models produced by PrefGeo retain more than **95%** of the performance of their corresponding single-preference models, while eliminating the substantial training costs of existing multi-preference alignment methods.

1 INTRODUCTION

Recent advances in text-to-image generation (Deng et al., 2025; Esser et al., 2024; Labs, 2024) have been driven by flow matching models (Kong et al., 2024; Polyak et al., 2024; Wan Team et al., 2025), achieving better visual quality. To align generated images with human preferences, post-training methods, particularly reinforcement learning from human feedback (RLHF) (Liu et al., 2026; Rafailov et al., 2023; Schulman et al., 2017; Xue et al., 2025), have become increasingly important. Reward-based RLHF typically learns reward models from human annotations to capture specific evaluation criteria, and optimizes the generator by sampling images, evaluating them with the reward models, and updating the model with reward-driven objectives.

However, human preferences for image quality are inherently multi-dimensional, encompassing criteria such as text-image alignment (Lin et al., 2024), aesthetic quality (Ke et al., 2023), text rendering (Cui et al., 2025), and preference (Ma et al., 2025). Despite this, existing methods are largely designed for single-preference optimization. Extending RL to multiple preferences is challenging: joint optimization can suffer from cross-preference interference, objective imbalance, catastrophic forgetting. Recent multi-preference alignment methods (Liu et al., 2026; Zheng et al., 2026; Chen et al., 2026; Li et al., 2026) have explored various strategies to address these challenges. However, they typically require additional alignment training, resulting in substantial GPU memory consumption and training time.

In this work, we propose PrefGeo, a training-free framework for multi-preference fusion. Building upon well-established single-preference alignment models, PrefGeo fuses the weights of multiple single-preference aligned models into a unified model, rather than jointly or sequentially training a single model for multiple preferences as in prior methods. By exploiting the preference geometric structure across preference-aligned models, PrefGeo eliminates the need for additional alignment training, thereby reducing training costs and the risk of cross-preference interference and catastrophic forgetting. This enables efficient multi-preference alignment with substantially lower GPU memory and running-time requirements.

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

PrefGeo first reveals a preference geometry showing that the weight matrices of preference-aligned experts can be approximately represented as left and right rotations of the corresponding weights of a shared base model. Building on this observation, PrefGeo consolidates multiple preference models by estimating and fusing their preference-specific rotations. Specifically, PrefGeo estimates these rotations by aligning the singular-vector bases of each expert with those of the shared base model via orthogonal Procrustes analysis (Schönemann, 1966). It then maps the rotations into skew-symmetric coordinates using the inverse Cayley transform (Macías-Virgós et al., 2018) and fuses them through norm-constrained aggregation to preserve their mean coordinate magnitude. Finally, the aggregated coordinates are mapped back to rotations and applied to the weights of base model, yielding a unified model without additional training or distillation. As shown in Figure 1, PrefGeo maintains individual preference performance across CLIP-Score (Hessel et al., 2021), GenEval (Ghosh et al., 2023), PickScore (Kirstain et al., 2023), and OCR (Chen et al., 2023), while substantially reducing GPU memory consumption and running time compared with state-of-the-art approaches such as DiffusionOPD (Li et al., 2026) and DiffusionNFT (Zheng et al., 2026).

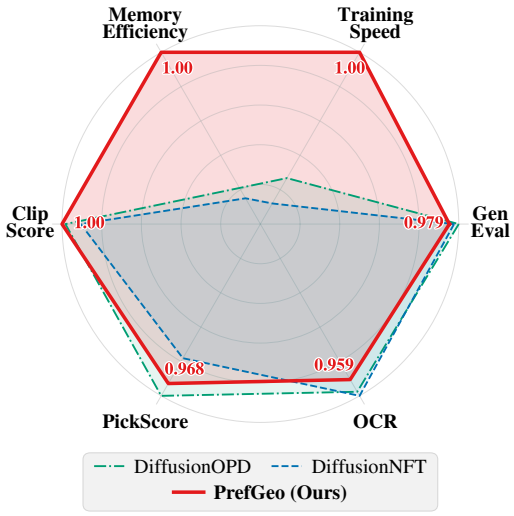


Figure 1: Comparison of multi-preference model construction efficiency and generation performance. We compare three methods, DiffusionOPD, DiffusionNFT, and PrefGeo, across six metrics: GenEval, OCR, PickScore, CLIPScore, Memory Efficiency, and Training Speed. Memory efficiency is defined as the reciprocal of peak GPU memory, while training speed is defined as the reciprocal of the running time required to obtain the multi-preference model. For each metric, the best result among the three methods is normalized to 1, and the other results are reported as percentages of the best value. Higher scores indicate better performance, and a larger covered area in the radar chart indicates stronger overall performance across the six metrics.

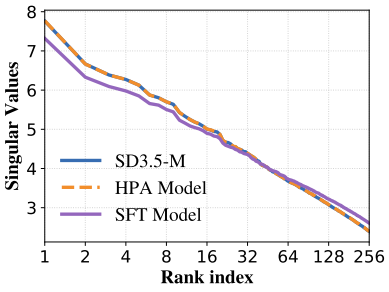
In summary, our contributions are summarized as follows:

- We reveal a preference geometry that connects the weights of preference-aligned experts to those of the shared base model through left and right rotations.
- We develop a multi-preference fusion method that obtains a unified model in only 18 minutes with 0.5GB of GPU memory, while preserving individual preferences.
- Extensive experiments demonstrate the effectiveness and flexibility of our method across diverse preferences and generation backbone.

2 Related Work

Our method, PrefGeo, formulates multi-preference alignment as a model fusion problem. In this section, we review existing approaches to multi-preference alignment and model fusion.

Multi-Preference Alignment integrates complementary capabilities from multiple specialized expert models into a unified model while preserving their individual strengths (Yadav et al., 2023; Zhang et al., 2025). However, existing approaches often rely on complex merging procedures, with additional computational costs. Specifically, MapReduce LoRA (Chen et al., 2026) trains preference-specific LoRA experts separately and integrates their capabilities through iterative merging, incurring the costs of both expert training and merging. DiffusionOPD (Li et al., 2026) consolidates multiple single-preference experts through on-policy distillation, but requires an additional student distillation stage, further increasing computational overhead. Therefore, efficiently consolidating existing preference experts without additional post-training or distillation remains a key challenge in the multi-preference alignment task.



(a)

Models	Weight Synthesis Methods	GenEval Score
HPA Model for GenEval	$\boldsymbol{W}_{\text{HPA}} \Rightarrow \boldsymbol{U}_{\text{HPA}}\boldsymbol{\Sigma}_{\text{HPA}}\boldsymbol{V}_{\text{HPA}}$	0.9204
	$\boldsymbol{W}_{\text{HPA}} \Leftarrow \boldsymbol{U}_{\text{HPA}}\boldsymbol{\Sigma}_{0}\boldsymbol{V}_{\text{HPA}}$	0.9224
	$\boldsymbol{W}_{\text{HPA}} \Leftarrow \boldsymbol{R}_{\text{L,HPA}}\boldsymbol{U}_{0}\boldsymbol{\Sigma}_{0}\boldsymbol{V}_{0}\boldsymbol{R}_{\text{R,HPA}}$	0.9224
SFT Model for GenEval	$\boldsymbol{W}_{\text{SFT}} \Rightarrow \boldsymbol{U}_{\text{SFT}}\boldsymbol{\Sigma}_{\text{SFT}}\boldsymbol{V}_{\text{SFT}}$	0.7512
	$\boldsymbol{W}_{\text{SFT}} \Leftarrow \boldsymbol{U}_{\text{SFT}}\boldsymbol{\Sigma}_{0}\boldsymbol{V}_{\text{SFT}}$	0.7286
	$\boldsymbol{W}_{\text{SFT}} \Leftarrow \boldsymbol{R}_{\text{L,SFT}}\boldsymbol{U}_{0}\boldsymbol{\Sigma}_{0}\boldsymbol{V}_{0}\boldsymbol{R}_{\text{R,SFT}}$	0.7286
SD3.5-M Model	$\boldsymbol{W}_{0} \Rightarrow \boldsymbol{U}_{0}\boldsymbol{\Sigma}_{0}\boldsymbol{V}_{0}$	0.6519

(b)

Figure 2: Evidence for preference geometry. (a) Top 256 singular values of the `Attn_to_k` weight matrix at layer 13. The human-preference-aligned (HPA) model based on SD3.5-M preserves the singular values of the base model, whereas supervised fine-tuning (SFT) starting from SD3.5-M induces deviations. (b) Performance variation under different parameter synthesis methods. The $\Rightarrow$ arrows denote SVD decomposition of $\boldsymbol{W}_{\text{HPA}}$, $\boldsymbol{W}_{\text{SFT}}$, and $\boldsymbol{W}_0$, while the $\Leftarrow$ arrows denote weight reconstruction using their SVD components and the corresponding rotation matrices. Replacing $\boldsymbol{\Sigma}_{\text{HPA}}$ with $\boldsymbol{\Sigma}_0$ leaves the GenEval score nearly unchanged, as shown in the second row. Similarly, reconstructing $\boldsymbol{W}_0$ using the left and right rotation matrices of the HPA model preserves its performance, as shown in the third row. In contrast, these reconstruction properties do not hold for the SFT model, supporting the existence of preference geometry in human preference alignment.

Model fusion merges the weights of multiple fine-tuned models to integrate their capabilities into a unified model without additional training (Jin et al., 2023; Ilharco et al., 2023; Yu et al., 2024). Task Arithmetic (Ilharco et al., 2023) represents parameter differences between fine-tuned models and the base model as task vectors and combines them through vector arithmetic. TSV-Merge (Gargiulo et al., 2025) uses SVD to obtain low-rank representations of task vectors, but does not model relationships among task-specific updates. OrthoMerge (Yang et al., 2026) explores model fusion from a geometric perspective by performing merging on the Riemannian manifold of the orthogonal group and extending the formulation to general fine-tuned models through Orthogonal-Residual Decoupling. However, its geometric formulation focuses on the intrinsic structure of model weights rather than the relationships among parameter updates from different preference experts. In contrast, PrefGeo explicitly characterizes these preference-specific relationships through a shared weight-space geometry and performs fusion in the resulting rotation space.

3 Phenomenon: The Geometry of Model Weights

Before developing our multi-preference fusion method, we first establish the empirical phenomenon that motivates our approach. PrefGeo is built upon a novel preference geometry that we discover and systematically characterize in this work. Specifically, we observe a consistent rotational relationship between the corresponding weights of a base model and its preference-aligned counterpart.

Preference Geometry. Let $\boldsymbol{\mathcal{M}}_0$ denote a pretrained flow matching model (Esser et al., 2024; Labs, 2024), serving as the base model, and let $\{\boldsymbol{\mathcal{M}}_i\}_{i=1}^{N}$ denote a set of preference-aligned expert models, each optimized for a specific human preference. Since the base model and each expert model share the same architecture, they have corresponding weight matrices at each layer ℓ. We denote the corresponding weight matrices at layer ℓ of the base model $\boldsymbol{\mathcal{M}}_0$ and the i-th expert model $\boldsymbol{\mathcal{M}}_i$ by $\boldsymbol{W}_0^\ell$ and $\boldsymbol{W}_i^\ell$, respectively. We then characterize their geometric relationship as follows:

Proposition 1. *Consider the weight matrices $\boldsymbol{W}_0^\ell$ and $\boldsymbol{W}_i^\ell$ at layer ℓ of the base model $\boldsymbol{\mathcal{M}}_0$ and the i-th preference-aligned model $\boldsymbol{\mathcal{M}}_i$. Let their singular value decompositions (SVDs) be*

$$\boldsymbol{W}_0^\ell = \boldsymbol{U}_0^\ell \boldsymbol{\Sigma}_0^\ell \boldsymbol{V}_0^{\ell\top}, \qquad \boldsymbol{W}_i^\ell = \boldsymbol{U}_i^\ell \boldsymbol{\Sigma}_i^\ell \boldsymbol{V}_i^{\ell\top}. \tag{1}$$

Empirically, we observe that preference alignment primarily induces rotations of the singular vector spaces while largely preserving the singular values. Specifically,

$$\boldsymbol{\Sigma}_i^\ell \approx \boldsymbol{\Sigma}_0^\ell, \qquad \boldsymbol{U}_i^\ell \approx \boldsymbol{R}_{\text{L},i}^\ell \boldsymbol{U}_0^\ell, \qquad \boldsymbol{V}_i^\ell \approx \boldsymbol{R}_{\text{R},i}^\ell \boldsymbol{V}_0^\ell. \tag{2}$$

where $\boldsymbol{R}_{\text{L},i}^\ell$ and $\boldsymbol{R}_{\text{R},i}^\ell$ are orthogonal transformations that characterize the preference-specific rotations of the left and right singular vector spaces, respectively.

Equation 2 introduces three approximate relationships that characterize the proposed preference geometry. In the following, we empirically validate each of these relationships.

Singular Value Invariance. We first examine whether preference alignment preserves the singular value of the model weights. We compare the singular values of $\boldsymbol{W}_0^\ell$ and $\boldsymbol{W}_i^\ell$ across ranks, with an SFT-trained model included as a reference. As shown in Figure 2(a), the singular value of the base model (SD3.5-M) and the human preference aligned model (HPA) for GenEval score remain highly consistent, in contrast to the SFT model. We further validate this observation from a performance perspective. As shown in Figure 2(b), replacing the HPA model's singular values with those of the base model preserves its specialized performance, supporting $\boldsymbol{\Sigma}_i^\ell \approx \boldsymbol{\Sigma}_0^\ell$ in Equation 2. In contrast, this reconstruction does not preserve the performance of the SFT model, which does not exhibit the singular value invariance for SFT model.

Left and Right Rotation Matrices. We next investigate the rotational relationships between the weight matrices of the preference-aligned models and the base model through left and right rotations. Specifically, we estimate $\boldsymbol{R}_{\mathrm{L},i}^\ell$ and $\boldsymbol{R}_{\mathrm{R},i}^\ell$, and measure the alignment of $\boldsymbol{U}_i^\ell$ with $\boldsymbol{R}_{\mathrm{L},i}^\ell \boldsymbol{U}_0^\ell$ and $\boldsymbol{V}_i^\ell$ with $\boldsymbol{R}_{\mathrm{R},i}^\ell \boldsymbol{V}_0^\ell$, respectively. These relationships can be indirectly validated by comparing the performance of the original expert models with that of models obtained by applying the estimated rotation matrices to the base model weights. As shown in Figure 2(b), the resulting models achieve performance comparable to the corresponding HPA models, supporting $\boldsymbol{U}_i^\ell \approx \boldsymbol{R}_{\mathrm{L},i}^\ell \boldsymbol{U}_0^\ell$ and $\boldsymbol{V}_i^\ell \approx \boldsymbol{R}_{\mathrm{R},i}^\ell \boldsymbol{V}_0^\ell$. In contrast, these relationships do not hold for the SFT model.

4 Our Method

We organize this section into three parts. First, we provide an overview of the overall framework. We then describe how preference-specific rotations are estimated from the aligned models, followed by the rotation fusion method that integrates them into a unified rotation.

Framework. The phenomenon observed in Section 3 reveals a geometric structure underlying preference alignment: the weight matrices $\boldsymbol{W}_i^\ell$ of a preference-aligned expert $\boldsymbol{\mathcal{M}}_i$ can be represented as rotated versions of the corresponding weight matrices $\boldsymbol{W}_0^\ell$ of the shared base model $\boldsymbol{\mathcal{M}}_0$. Specifically, the empirical relationships in Equation 2 yield the following weight-space transformation:

$$\boldsymbol{W}_i^\ell \approx \boldsymbol{R}_{\mathrm{L},i}^\ell \boldsymbol{W}_0^\ell \boldsymbol{R}_{\mathrm{R},i}^{\ell\top}. \tag{3}$$

This provides a geometric representation of each preference-aligned expert through its preference-specific rotations $\{\boldsymbol{R}_{\mathrm{L},i}^\ell, \boldsymbol{R}_{\mathrm{R},i}^\ell\}_{i=1}^N$, motivating PrefGeo to perform multi-preference fusion in the rotation space. Specifically, PrefGeo fuses the rotations associated with different preferences into a pair of transformations $\{\boldsymbol{R}_{\mathrm{L}}^\ell, \boldsymbol{R}_{\mathrm{R}}^\ell\}$. The resulting multi-preference fused weight matrix is given by

$$\boldsymbol{W}^\ell \approx \boldsymbol{R}_{\mathrm{L}}^\ell \boldsymbol{W}_0^\ell \boldsymbol{R}_{\mathrm{R}}^{\ell\top}. \tag{4}$$

Applying these fused transformations to the shared base model yields a model that jointly captures the multiple preferences encoded by the individual experts, while maintaining minimal degradation in individual preference performance and requiring no additional preference alignment training. Following this formulation, we develop two key procedures in the subsequent paragraphs. First, we estimate the preference-specific rotation matrices $\{\boldsymbol{R}_{\mathrm{L},i}^\ell, \boldsymbol{R}_{\mathrm{R},i}^\ell\}_{i=1}^N$ for each preference-aligned expert. Second, we fuse the estimated rotations across multiple experts into a unified pair of transformations $\{\boldsymbol{R}_{\mathrm{L}}^\ell, \boldsymbol{R}_{\mathrm{R}}^\ell\}$ to construct the final multi-preference model.

Rotation Estimation. As observed in Equation 2, each preference-aligned expert induces left and right rotation matrices relative to the base model. We estimate these rotation matrices by solving Orthogonal Procrustes problem (Schönemann, 1966):

$$\begin{aligned} &\boldsymbol{R}_{\mathrm{L},i}^\ell = \arg\min_{\boldsymbol{R}_{\mathrm{L}}} \left\| \boldsymbol{U}_i^\ell - \boldsymbol{R}_{\mathrm{L}} \boldsymbol{U}_0^\ell \right\|_F^2, \qquad \boldsymbol{R}_{\mathrm{R},i}^\ell = \arg\min_{\boldsymbol{R}_{\mathrm{R}}} \left\| \boldsymbol{V}_i^\ell - \boldsymbol{R}_{\mathrm{R}} \boldsymbol{V}_0^\ell \right\|_F^2, \\ &\text{s.t.} \quad \boldsymbol{R}_{\mathrm{L}}^\top \boldsymbol{R}_{\mathrm{L}} = \boldsymbol{I},\ \det(\boldsymbol{R}_{\mathrm{L}}) = 1. \qquad \text{s.t.} \quad \boldsymbol{R}_{\mathrm{R}}^\top \boldsymbol{R}_{\mathrm{R}} = \boldsymbol{I},\ \det(\boldsymbol{R}_{\mathrm{R}}) = 1. \end{aligned} \tag{5}$$

The closed-form solution (Gower & Dijksterhuis, 2004) is given by:

$$\boldsymbol{R}_{\mathrm{L},i}^\ell = \boldsymbol{U}_i^\ell \boldsymbol{U}_0^{\ell\top}, \qquad \boldsymbol{R}_{\mathrm{R},i}^\ell = \boldsymbol{V}_i^\ell \boldsymbol{V}_0^{\ell\top}. \tag{6}$$

Rotation Fusion. Directly fusing preference-specific rotation matrices is non-trivial: although each rotation matrix is orthogonal, their weighted average is generally no longer orthogonal. To enable valid rotation fusion, we map the rotations to a Cayley representation space (Macías-Virgós et al., 2018) in which linear fusion preserves the desired structure. Specifically, for an orthogonal matrix $\boldsymbol{R}$ and a skew-symmetric matrix $\boldsymbol{Q}$, the inverse Cayley transform $\mathfrak{C}^{-1}(\boldsymbol{R})$ and Cayley transform $\mathfrak{C}(\boldsymbol{Q})$ are respectively defined as

$$\boldsymbol{Q} = \mathfrak{C}^{-1}(\boldsymbol{R}) = (\boldsymbol{R} - \boldsymbol{I})(\boldsymbol{R} + \boldsymbol{I})^{-1}. \tag{7}$$

$$\boldsymbol{R} = \mathfrak{C}(\boldsymbol{Q}) = (\boldsymbol{I} + \boldsymbol{Q})(\boldsymbol{I} - \boldsymbol{Q})^{-1}. \tag{8}$$

These transforms map rotation matrices to skew-symmetric matrices, enabling rotation fusion in the skew-symmetric space. Since each preference-specific rotation consists of a pair of transformations, $\{\boldsymbol{R}^{\ell}_{\mathrm{L},i}\}_{i=1}^{N}$ and $\{\boldsymbol{R}^{\ell}_{\mathrm{R},i}\}_{i=1}^{N}$, we present the fusion of the left rotation $\{\boldsymbol{R}^{\ell}_{\mathrm{L},i}\}_{i=1}^{N}$ procedure below for brevity; the same procedure applies to the right rotation $\{\boldsymbol{R}^{\ell}_{\mathrm{R},i}\}_{i=1}^{N}$.

Let $\boldsymbol{Q}^{\ell}_{\mathrm{L},i}$ denote the skew-symmetric representation of the i-th left rotation, obtained by

$$\boldsymbol{Q}^{\ell}_{\mathrm{L},i} = \mathfrak{C}^{-1}(\boldsymbol{R}^{\ell}_{\mathrm{L},i}). \tag{9}$$

PrefGeo formulates rotation fusion as minimizing the sum of squared Frobenius distances between the fused skew-symmetric representation $\boldsymbol{Q}^{\ell}$ and the preference-specific representations $\{\boldsymbol{Q}^{\ell}_{\mathrm{L},i}\}_{i=1}^{N}$:

$$\begin{aligned} &\min_{\boldsymbol{Q}^{\ell}} \sum_{i=1}^{N} \|\boldsymbol{Q}^{\ell} - \boldsymbol{Q}^{\ell}_{\mathrm{L},i}\|_F^2 \\ &\text{s.t.} \quad \|\boldsymbol{Q}^{\ell}\|_F = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{Q}^{\ell}_{\mathrm{L},i}\|_F. \end{aligned} \tag{10}$$

We note that the constraint preserves the average magnitude of the preference-specific rotations and the problem admits the following closed-form global minimizer (Proof in Appendix A):

$$\boldsymbol{Q}^{\ell} = \frac{\sum_{i=1}^{N} \|\boldsymbol{Q}^{\ell}_{\mathrm{L},i}\|_F}{N \cdot \left\|\sum_{i=1}^{N} \boldsymbol{Q}^{\ell}_{\mathrm{L},i}\right\|_F} \sum_{i=1}^{N} \boldsymbol{Q}^{\ell}_{\mathrm{L},i}. \tag{11}$$

Note that any linear combination of skew-symmetric matrices remains skew-symmetric. We can therefore apply the Cayley transform to $\boldsymbol{Q}^{\ell}$ to obtain the fused left rotation:

$$\boldsymbol{R}^{\ell}_{\mathrm{L}} = \mathfrak{C}(\boldsymbol{Q}^{\ell}). \tag{12}$$

Moreover, applying the above procedure to $\{\boldsymbol{R}^{\ell}_{\mathrm{R},i}\}_{i=1}^{N}$ yields the fused right rotation $\boldsymbol{R}^{\ell}_{\mathrm{R}}$. By estimating and fusing preference-specific rotations, PrefGeo performs multi-preference fusion without jointly optimizing competing preference objectives. This design mitigates common issues in training-based multi-preference alignment, such as cross-preference interference, objective imbalance, and catastrophic forgetting. At the same time, PrefGeo substantially reduces the computational cost of model construction while maintaining the performance of individual preferences.

5 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the effectiveness and flexibility of our method and provide further empirical evidence supporting the existence of preference geometry in preference-aligned models.

5.1 EXPERIMENTAL SETUP

Benchmarks and Evaluation Metrics. We evaluate our method on three tasks: compositional fidelity, visual text rendering, and aesthetic preference. For compositional fidelity and visual text rendering, we use GenEval (Ghosh et al., 2023) and OCR (Chen et al., 2023), respectively, adopting the evaluation splits provided by FlowGRPO (Liu et al., 2026). For aesthetic preference (AES), we evaluate on DrawBench (Saharia et al., 2022) using PickScore (Kirstain et al., 2023), CLIPScore (Hessel et al., 2021), HPSv2.1 (Wu et al., 2023), ImageReward (Xu et al., 2023), and UnifiedReward (Wang et al., 2025).

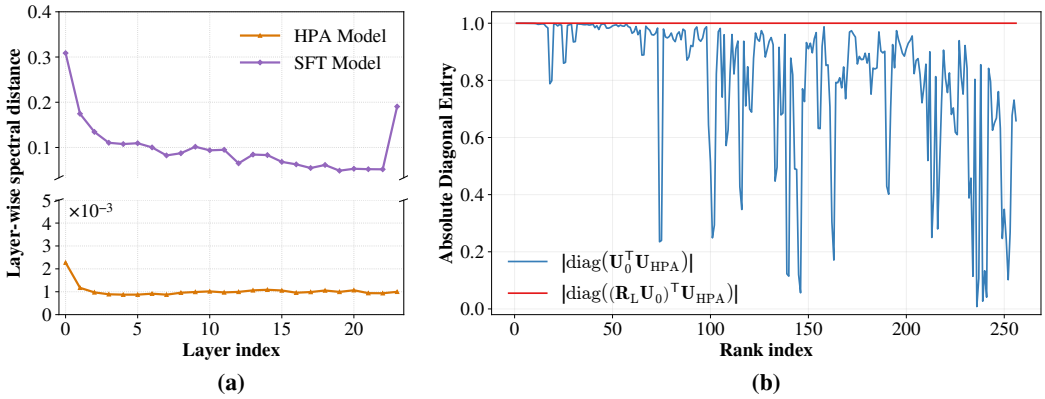


Figure 3: Further Evidence for Preference Geometry. (a) Broken-axis plot of the layer-wise spectral distances between the base model (SD3.5-M) and its aligned counterparts. Each point represents the normalized spectral distance $d_{\mathbf{\Sigma}}$ 13 computed from the `Attn_to_k` weight matrices of each layer. The human-preference-aligned (HPA) model exhibits consistently small spectral deviations across layers, whereas the SFT model induces substantially larger changes. (b) Illustration for absolute diagonal entries $|\operatorname{diag}(\boldsymbol{U}_0^\top \boldsymbol{U}_{\mathrm{HPA}})|$ and $|\operatorname{diag}((\boldsymbol{R}_{\mathrm{L}}\boldsymbol{U}_0)^\top \boldsymbol{U}_{\mathrm{HPA}})|$ across singular-vector ranks. The near-unit diagonal entries of $(\boldsymbol{R}_{\mathrm{L}}\boldsymbol{U}_0)^\top \boldsymbol{U}_{\mathrm{HPA}}$ indicate that the extracted rotation $\boldsymbol{R}_{\mathrm{L}}$ accurately captures the directional changes in the left singular vectors induced by preference alignment.

Implementation Details. We use SD3.5-M (Esser et al., 2024) as the base model and conduct all experiments on a single NVIDIA RTX 4090 24GB GPU. Unless otherwise specified, PrefGeo fuses three single-preference-aligned experts corresponding to GenEval, OCR, and AES. At inference, we use 40 steps ODE sampler with CFG $= 4.0$, and generate image with 512×512 resolution. To demonstrate the flexibility of PrefGeo, we further extend our method to FLUX.1-Dev (Labs, 2024). (More results on FLUX.1-Dev are provided in Appendix B)

5.2 More Evidence of Preference Geometry

We conduct additional experiments to validate the phenomenon of preference geometry. As shown in Figure 3(a), we quantify the layer-wise spectral differences between the base model and its aligned counterparts using the normalized spectral distance:

$$d_{\mathbf{\Sigma}} = \frac{\|\mathbf{\Sigma}_i^\ell - \mathbf{\Sigma}_0^\ell\|_F}{\|\mathbf{\Sigma}_0^\ell\|_F + \varepsilon}. \tag{13}$$

where ε is a small constant introduced to avoid numerical instability when $|\mathbf{\Sigma}_0^\ell|_F$ is close to zero. The spectral distances between the human-preference-aligned (HPA) model and the base model (SD3.5-M) remain below 2×10^{-3} across all layers, whereas those of the supervised fine-tuning (SFT) model consistently exceed 0.05. These results provide empirical evidence for singular value invariance in preference-aligned models. As shown in Figure 3(b), we compare the absolute diagonal entries $|\operatorname{diag}(\boldsymbol{U}_0^\top \boldsymbol{U}_{\mathrm{HPA}})|$ and $|\operatorname{diag}((\boldsymbol{R}_{\mathrm{L}}\boldsymbol{U}_0)^\top \boldsymbol{U}_{\mathrm{HPA}})|$. After applying the rotation matrices, the diagonal entries approach unity across singular vector ranks, indicating that $\boldsymbol{R}_{\mathrm{L}}$ accurately captures the directional changes induced by preference alignment. Although only the left rotation is shown, the same observation holds for the right rotation. Together, these results provide further empirical evidence supporting the proposed preference geometry.

5.3 Performance Comparison

We compare **PrefGeo** against several competitive baselines: (1) **single-preference-aligned models**, using the task-specific GenEval, OCR, and aesthetics (AES) models provided by DiffusionOPD (Li et al., 2026); (2) **multi-preference-aligned models**, including DiffusionOPD, DiffusionNFT (Zheng et al., 2026), MapReduce LoRA (Chen et al., 2026), and Multi-task GRPO (Liu et al., 2026); and (3) **model fusion methods** originally designed for large language models (LLMs), including Task Arithmetic (TA) (Ilharco et al., 2023), TSV-M (Gargiulo et al., 2025), and TIES (Yadav et al., 2023). We adapt these methods to SD3.5-M for experiments.



Figure 4: Visual comparison of PrefGeo with single and multi preference alignment methods. PrefGeo consistently produces images with high fidelity and strong structural coherence. Red dashed boxes highlight generation errors. The same random seed is used across all methods for each prompt.

5.3.1 Visual Comparison

As shown in Figure 4, we provide visual comparison of PrefGeo with 3 single-preference-aligned models and 3 multi-preference-aligned models. PrefGeo effectively integrates the capabilities of the three experts, resulting in images with accurate object counts, correct attribute binding, faithful text rendering, accurate spatial relationships, vibrant colors, and fine-grained details. Furthermore, we compare PrefGeo with three model fusion methods by adapting Task Arithmetic (TA), TSV-M, and TIES, which were originally designed for LLMs, to SD3.5-M. As shown in Figure 5, directly applying these conventional model fusion methods to flow-matching models often results in noticeable degradation in image quality. Employing preference geometry, PrefGeo effectively extends the model fusion paradigm to flow-matching models while maintaining high-quality image generation.

5.3.2 Quantitative Results

PrefGeo achieves a trade-off between generation performance and computational efficiency. Table 1 presents the quantitative comparison of PrefGeo with several baselines. By fusing existing single-preference experts, PrefGeo requires only 0.3 h, compared with GRPO-Mix (1973.6 h), DiffusionNFT (1952.7 h), and DiffusionOPD (171.2 h). PrefGeo achieves a 570× speedup over DiffusionOPD in the reported wall-clock time. Beyond its low computational cost, PrefGeo also achieves an average score of 0.871, demonstrating generation performance comparable to the state-of-the-art DiffusionOPD (0.894), while outperforming other multi-preference alignment methods, including

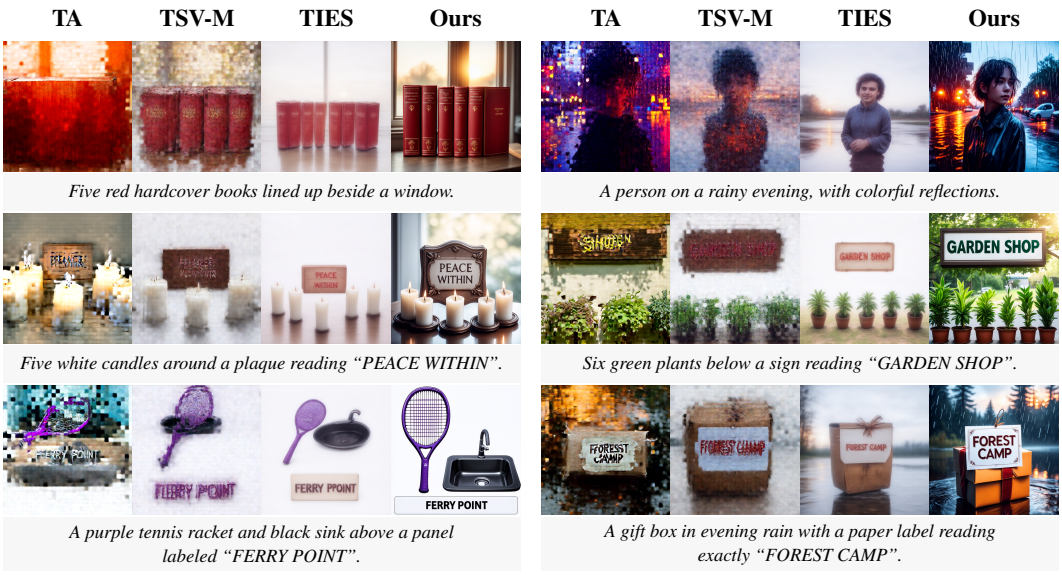


Figure 5: Visual comparison of PrefGeo against model-merging methods. PrefGeo effectively satisfies multiple preference objectives within a single prompt.

GRPO-Mix (0.857) and DiffusionNFT (0.830). Moreover, PrefGeo outperforms all model merging methods, including Task Arithmetic (0.532), TSV-M (0.543) and TIES (0.669). Specifically, PrefGeo retains over 95% of the performance of its corresponding preference experts across all in-domain metrics while achieving competitive results on out-of-domain metrics. These results demonstrate that PrefGeo achieves strong multi-preference alignment capabilities. Furthermore, as shown in the last row of Table 1, PrefGeo can be extended to FLUX.1-Dev. These results highlight the flexibility of our method across different generation backbones.

Table 1: Evaluation Results. Gray-colored: Base model. Green-colored: Our method. **Bold**: best, Underline: second best. ↑: higher is better. Average column reports the mean score across all metrics after min-max normalization. Wall-clock time (hours) denotes training or fusion time.

Model	In-Domain					Out-of-Domain		Average ↑	Wall-clock time (h)
	GenEval ↑	OCR ↑	PickScore ↑	ClipScore ↑	HPSv2.1 ↑	ImgRwd ↑	UniRwd ↑		
SD3.5-M	0.65	0.69	21.69	0.284	0.284	0.85	3.02	0.702	–
+GenEval Teacher	0.92	0.79	20.99	0.287	0.267	0.87	3.17	0.751	713.4
+OCR Teacher	0.63	**0.97**	21.38	0.291	0.279	0.93	3.06	0.746	504.3
+Aes Teacher	0.72	0.87	**23.69**	0.297	**0.345**	**1.49**	**3.48**	0.849	1303.8
GRPO-Mix	0.89	**0.97**	22.90	0.296	0.316	1.34	3.37	0.857	1973.6
DiffusionNFT	0.93	**0.97**	22.06	0.294	0.291	1.18	3.29	0.830	1952.7
DiffusionOPD	**0.94**	**0.97**	23.67	0.298	0.339	**1.49**	3.47	**0.894**	171.2
MapReduce LoRA	0.87	0.95	22.50	0.293	0.299	1.21	3.32	0.827	1711.1
Task Arithmetic	0.41	0.47	20.20	0.244	0.221	0.579	2.01	0.532	0.4
TSV-M	0.41	0.44	20.03	0.244	0.227	0.54	2.09	0.543	0.6
TIES	0.64	0.65	21.55	0.278	0.282	0.815	2.93	0.669	0.4
PrefGeo (Ours)	0.92	0.93	22.92	**0.299**	0.328	1.46	3.45	0.871	0.3
FLUX.1-Dev	0.64	0.66	21.99	0.272	0.302	0.91	3.20	0.719	–
MapReduce LoRA	**0.87**	**0.97**	23.03	**0.318**	0.323	1.34	**3.51**	**0.883**	6035.7
PrefGeo (Ours)	0.86	**0.97**	**23.33**	0.316	**0.329**	**1.35**	3.45	0.881	1.4

5.3.3 User Study

To evaluate the effectiveness of PrefGeo from a human perspective, we select 20 representative prompts from GenEval, OCR, and DrawBench, respectively. For each prompt, we compare PrefGeo against five baselines based on SD3.5-M. We recruit 30 participants who independently rank the generated images from best to worst. Then we assign scores from 6 to 1 based on these rankings. As shown in Figure 6, our method achieves comparable performance with the SOTA method and single preference aligned model, demonstrating the effectiveness of the fusion.

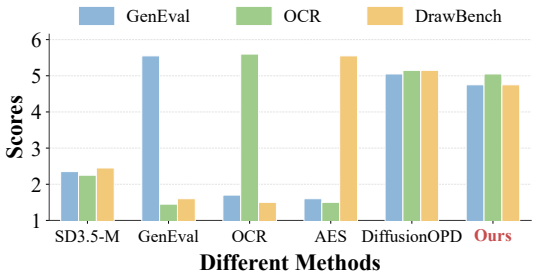


Figure 6: User study results. Higher scores indicate better human performance.

Table 2: Alignment performance with different numbers of fused single-preference-aligned models. GE denotes GenEval model.

Experts	GenEval	OCR	PickScore
GE	0.92	0.80	20.99
OCR	0.63	0.97	21.38
AES	0.71	0.86	23.69
GE + OCR	0.91	0.93	21.28
GE + AES	0.93	0.86	22.80
OCR + AES	0.73	0.95	23.18
GE + OCR + AES	0.92	0.93	22.92

5.4 ABLATION STUDY

In this section, we conduct ablation studies to investigate the design choices of our method. All experiments follow the settings described in the Implementation Details section.

Number of Single-Preference-Aligned Models. PrefGeo fuses multiple single-preference-aligned models into a unified model. We investigate how the number of preference-aligned models affects the performance of the fused model across different preference metrics. As shown in Table 2, fusing two experts achieves strong performance on their corresponding preference metrics. When all three experts are fused, PrefGeo maintains balanced performance across GenEval (0.92), OCR (0.93), and PickScore (22.92), demonstrating its ability to integrate an increasing number of preferences. We consider up to three experts because only three publicly available preference-aligned models with distinct preference objectives are currently available in the existing literature.

Fusion Space. To validate the necessity of the Cayley transform in PrefGeo, we compare fusion in different representation spaces. As shown in Table 3, directly fusing weights in the weight space ($\boldsymbol{W}$) via weighted averaging achieves reasonable performance but results in substantial performance degradation. PrefGeo instead extracts preference-specific rotation matrices from the expert models. However, directly averaging these rotation matrices ($\boldsymbol{R}$) generally violates orthogonality, leading to severe performance degradation across all metrics. By performing fusion in the Cayley representation space ($\boldsymbol{Q}$), PrefGeo preserves the orthogonality of the fused rotations and achieves the highest performance on GenEval, OCR, and PickScore.

Table 3: Alignment performance with different fusion space. $\boldsymbol{W}$ Space: Weight Space; $\boldsymbol{R}$ Space: Rotation Space; $\boldsymbol{Q}$ Space: Cayley Space.

Fusion Space	GenEval	OCR	PickScore
$\boldsymbol{W}$ Space	0.85	0.89	22.39
$\boldsymbol{R}$ Space	0.46	0.32	20.23
$\boldsymbol{Q}$ Space (Ours)	**0.92**	**0.93**	**22.92**

6 CONCLUSION

In this paper, we introduce PrefGeo, a model fusion framework for efficient multi-preference alignment in flow-matching-based text-to-image generation. Our framework is motivated by an empirical observation that preference alignment approximately preserves the singular spectrum of the base model while modifying its singular-vector directions, revealing a rotational relationship between preference-aligned models and their shared base model. Building on this preference geometry, PrefGeo fuses multiple single-preference experts through geometric operations on model weights without additional training. Extensive experiments demonstrate that PrefGeo effectively integrates multiple preferences while retaining over 95% of the performance of the corresponding single-preference experts. Moreover, PrefGeo completes model fusion in only 18 minutes on a single NVIDIA RTX 4090 24GB GPU, using just 0.5GB of GPU memory. These results highlight the potential of preference geometry as an efficient alternative to training-based multi-preference alignment.

AI USE STATEMENT

In this work, we used generative AI tools for [provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results], and [Generate synthetic data sets, clean and reformat dataset] are not applicable to this work. Additionally, we used generative AI tools for [Formulate questions for surveys or interviews, create or modify scientific figures or images, create or edit software code, sourcing/searching for information, edit a research paper to improve readability, propose a title or keywords for a research paper]. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies multi-preference alignment for text-to-image generation and does not introduce or collect sensitive personal data. Our human evaluation involves participants who assess generated images solely for research purposes. Participation is voluntary, and the evaluation does not require participants to provide personally identifiable or sensitive information. The prompts and generated images used in the evaluation are selected to avoid harmful, offensive, or privacy-sensitive content. As with other text-to-image generation methods, the proposed approach may potentially be misused to generate misleading or inappropriate content. We encourage responsible use of the method and compliance with applicable policies and regulations.

REPRODUCIBILITY STATEMENT

We use publicly available datasets and open-source models in our experiments. We provide detailed descriptions of the proposed method, implementation details, and experimental settings to facilitate reproducibility.

REFERENCES

Chieh-Yun Chen, Zhonghao Wang, Qi Chen, Zhifan Ye, Min Shi, Yue Zhao, Yinan Zhao, Hui Qu, Wei-An Lin, Yiru Shen, et al. MapReduce LoRA: Advancing The Pareto Front in Multi-Preference Optimization for Generative Models. In *Conference on Computer Vision and Pattern Recognition*, 2026.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. TextDiffuser: Diffusion Models as Text Painters. In *Advances in Neural Information Processing Systems*, 2023.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. PaddleOCR 3.0 Technical Report. *ArXiv*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *ArXiv*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *International Conference on Machine Learning*, 2024.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task Singular Vectors: Reducing Task Interference in Model Merging. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. In *Advances in Neural Information Processing Systems*, 2023.

John C Gower and Garmt B Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-Free Evaluation Metric for Image Captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations*, 2023.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless Knowledge Fusion by Merging Weights of Language Models. In *International Conference on Learning Representations*, 2023.

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In *Conference on Computer Vision and Pattern Recognition*, 2023.

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. In *Advances in Neural Information Processing Systems*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *ArXiv*, 2024.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. DiffusionOPD: A Unified Perspective of On-Policy Distillation in Diffusion Models. In *SIGGRAPH Asia 2026 Conference Papers*, 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In *European Conference on Computer Vision*, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. In *Advances in Neural Information Processing Systems*, 2026.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *International Conference on Computer Vision*, 2025.

Enrique Macías-Virgós, María José Pereira-Sáez, and Daniel Tanré. Cayley transform on stiefel manifolds. *Journal of Geometry and Physics*, 2018.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix

Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie Gen: A Cast of Media Foundation Models. *ArXiv*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *ArXiv*, 2017.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 1966.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *ArXiv*, 2025.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *ArXiv*, 2025.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score: Better Aligning Text-to-Image Models with Human Preference. In *International Conference on Computer Vision*, 2023.

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. In *Advances in Neural Information Processing Systems*, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on Visual Generation. *ArXiv*, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In *Advances in Neural Information Processing Systems*, 2023.

Sihan Yang, Kexuan Shi, and Weiyang Liu. Orthogonal Model Merging. In *International Conference on Machine Learning*, 2026.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models Are Super Mario: Absorbing Abilities from Homologous Models as A Free Lunch. In *International Conference on Machine Learning*, 2024.

Mingyang Zhang, Jing Liu, Ganggui Ding, Linlin Ou, Xinyi Yu, and Bohan Zhuang. Channel Merging: Preserving Specialization for Merged Experts. *Conference on Artificial Intelligence*, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026.

A Derivation of the Closed-Form Rotation Fusion Solution

In this section, we derive the closed-form solution to the rotation fusion problem. We focus on the left rotation Matrices $\boldsymbol{R}_{\mathrm{L},i}^{\ell}$at layer ℓ; the same derivation applies to the right rotation Matrices $\boldsymbol{R}_{\mathrm{R},i}^{\ell}$.

The preference-specific representations $\{\boldsymbol{Q}_{\mathrm{L},i}^{\ell}\}_{i=1}^{N}$ are obtained through the inverse Cayley transform:

$$\boldsymbol{Q}_{\mathrm{L},i}^{\ell} = \mathfrak{C}^{-1}(\boldsymbol{R}_{\mathrm{L},i}^{\ell}) = (\boldsymbol{R}_{\mathrm{L},i}^{\ell} - \boldsymbol{I})(\boldsymbol{R}_{\mathrm{L},i}^{\ell} + \boldsymbol{I})^{-1}. \tag{14}$$

We assume that each $\boldsymbol{R}_{\mathrm{L},i}^{\ell} + \boldsymbol{I}$ is invertible, so that the inverse Cayley transform is well-defined. Since the input rotations are orthogonal, their Cayley representations satisfy $(\boldsymbol{Q}_{\mathrm{L},i}^{\ell})^{\top} = -\boldsymbol{Q}_{\mathrm{L},i}^{\ell}$.

The fusion problem is

$$\begin{aligned} \min_{\boldsymbol{Q}^{\ell}} \quad & \sum_{i=1}^{N} \|\boldsymbol{Q}^{\ell} - \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F^2, \\ \text{s.t.} \quad & \|\boldsymbol{Q}^{\ell}\|_F = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F. \end{aligned} \tag{15}$$

We first solve this problem over real matrices of the same size as the input representations, and then verify that its solution is skew-symmetric.

Expansion of the objective. Expanding the squared Frobenius distances gives

$$\begin{aligned} &\sum_{i=1}^{N} \|\boldsymbol{Q}^{\ell} - \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F^2 \\ &\quad = N\|\boldsymbol{Q}^{\ell}\|_F^2 + \sum_{i=1}^{N} \|\boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F^2 - 2\,\mathrm{tr}\left((\boldsymbol{Q}^{\ell})^{\top} \sum_{i=1}^{N} \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\right). \end{aligned} \tag{16}$$

Under the norm constraint, this becomes

$$\begin{aligned} &\sum_{i=1}^{N} \|\boldsymbol{Q}^{\ell} - \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F^2 \\ &\quad = \frac{1}{N}\left(\sum_{i=1}^{N} \|\boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F\right)^2 + \sum_{i=1}^{N} \|\boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F^2 \\ &\qquad - 2\,\mathrm{tr}\left((\boldsymbol{Q}^{\ell})^{\top} \sum_{i=1}^{N} \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\right). \end{aligned} \tag{17}$$

The first two terms are independent of $\boldsymbol{Q}^{\ell}$. Consequently, minimizing the objective is equivalent to maximizing

$$\mathrm{tr}\left((\boldsymbol{Q}^{\ell})^{\top} \sum_{i=1}^{N} \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\right) \tag{18}$$

under the same norm constraint.

Global optimality and uniqueness. Suppose that $\sum_{i=1}^{N} \boldsymbol{Q}_{\mathrm{L},i}^{\ell} \neq \boldsymbol{0}$. By the Cauchy–Schwarz inequality for the Frobenius inner product,

$$\begin{aligned} \mathrm{tr}\left((\boldsymbol{Q}^{\ell})^{\top} \sum_{i=1}^{N} \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\right) &\leq \|\boldsymbol{Q}^{\ell}\|_F \left\|\sum_{i=1}^{N} \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\right\|_F \\ &= \frac{1}{N}\left(\sum_{i=1}^{N} \|\boldsymbol{Q}_{\mathrm{L},i}^{\ell}\|_F\right) \left\|\sum_{i=1}^{N} \boldsymbol{Q}_{\mathrm{L},i}^{\ell}\right\|_F. \end{aligned} \tag{19}$$

Because the sum of the input representations is nonzero, the prescribed norm is strictly positive. Equality in equation 19 holds if and only if $\boldsymbol{Q}^\ell$ is a positive scalar multiple of $\sum_{i=1}^N \boldsymbol{Q}^\ell_{\mathrm{L},i}$. The norm constraint uniquely determines the scaling, yielding equation 11:

$$\boxed{\boldsymbol{Q}^\ell = \frac{\sum_{i=1}^N \|\boldsymbol{Q}^\ell_{\mathrm{L},i}\|_F}{N \left\|\sum_{i=1}^N \boldsymbol{Q}^\ell_{\mathrm{L},i}\right\|_F} \sum_{i=1}^N \boldsymbol{Q}^\ell_{\mathrm{L},i}.} \tag{20}$$

This matrix satisfies the norm constraint, since

$$\begin{aligned}\|\boldsymbol{Q}^\ell\|_F &= \frac{\sum_{i=1}^N \|\boldsymbol{Q}^\ell_{\mathrm{L},i}\|_F}{N \left\|\sum_{i=1}^N \boldsymbol{Q}^\ell_{\mathrm{L},i}\right\|_F} \left\|\sum_{i=1}^N \boldsymbol{Q}^\ell_{\mathrm{L},i}\right\|_F \\ &= \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{Q}^\ell_{\mathrm{L},i}\|_F. \end{aligned} \tag{21}$$

It also attains the upper bound in equation 19, and therefore minimizes the original objective globally. The equality condition and the fixed norm establish that this global minimizer is unique.

Preservation of skew-symmetry. Since each input representation is skew-symmetric,

$$\left(\sum_{i=1}^N \boldsymbol{Q}^\ell_{\mathrm{L},i}\right)^\top = -\sum_{i=1}^N \boldsymbol{Q}^\ell_{\mathrm{L},i}. \tag{22}$$

The solution in equation 20 is a real scalar multiple of this sum and hence satisfies

$$(\boldsymbol{Q}^\ell)^\top = -\boldsymbol{Q}^\ell. \tag{23}$$

Thus, the unique global minimizer also belongs to the skew-symmetric representation space. In particular, explicitly imposing skew-symmetry in equation 15 does not change the solution in the nondegenerate case. The fused left rotation is then obtained as

$$\boldsymbol{R}^\ell_{\mathrm{L}} = \mathfrak{C}(\boldsymbol{Q}^\ell) = (\boldsymbol{I} + \boldsymbol{Q}^\ell)(\boldsymbol{I} - \boldsymbol{Q}^\ell)^{-1}. \tag{24}$$

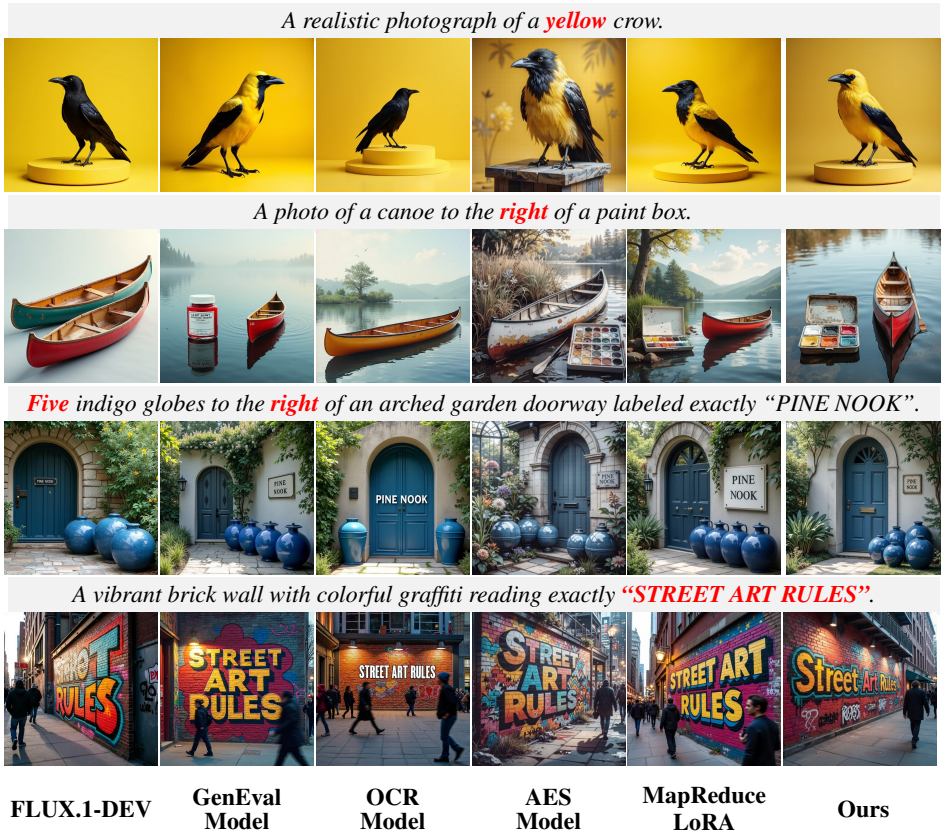


Figure 7: Visual comparison of PrefGeo with MapReduce LoRA and single-preference experts based on FLUX.1-Dev.

B Evaluation of Flexibility Across different Base Models

To further evaluate the flexibility of PrefGeo across different base models, we extend our method from SD3.5-M to FLUX.1-Dev. We report both qualitative and quantitative results to demonstrate its applicability under a different model backbone.

Implementation Details. All experiments are conducted on a single NVIDIA RTX 4090 24GB GPU. We use FLUX.1-Dev as the base model. At inference, we use 28 sampling steps with a guidance scale of 3.5 and generate images at a resolution of 512×512. All quantitative and qualitative results are obtained using the same random seed for a fair comparison.

Qualitative Results. Figure 7 presents qualitative comparisons based on FLUX.1-Dev. PrefGeo effectively integrates the capabilities of different preference experts into a single model. These results show that the multi-preference fusion ability of PrefGeo is not limited to the base model used in our main experiments and can be naturally transferred to a different backbone.

Quantitative Results. Table 4 presents the quantitative results on FLUX.1-Dev. PrefGeo maintains competitive performance across both in-domain and out-of-domain evaluations, indicating that its multi-preference fusion remains effective when applied to a different base model.

Table 4: Evaluation Results on FLUX.1-Dev. All methods are evaluated under 512×512 resolution with the same random seed. Gray-colored: Base model. Green-colored: Our method. **Bold**: best, Underline: second best. Average column reports the mean score across all metrics after min-max normalization.

Model	In-Domain					Out-of-Domain		Average
	GenEval	OCR	PickScore	ClipScore	HPSv2.1	ImgRwd	UniRwd	
FLUX.1-Dev	0.64	0.66	21.99	0.272	0.302	0.91	3.20	0.719
+GenEval Teacher	**0.88**	0.79	22.09	0.311	0.309	1.13	3.40	0.820
+OCR Teacher	0.67	**0.98**	22.03	0.311	0.306	1.05	3.29	0.804
+Aes Teacher	0.66	0.88	**23.58**	0.310	0.326	1.33	3.46	0.837
MapReduce LoRA	0.87	0.97	23.03	**0.318**	0.323	1.34	**3.51**	**0.883**
Ours	0.86	0.97	23.33	0.316	**0.329**	**1.35**	3.45	0.881

C MORE VISUALIZATION RESULTS

We provide additional sample images based on SD3.5-M as shown in Figure 8, Figure 9. And sample images based on FLUX.1-Dev as shown in Figure 10, Figure 11

Table 5: Detailed prompts used to generate the images in Figure 8.

Image	Prompt
Row 1, Col 1	An oil painting of astronauts in space.
Row 1, Col 2	A close-up portrait of a girl with an autumn leaves headdress.
Row 1, Col 3	A skyline with several buildings with old architecture.
Row 1, Col 4	A futuristic building towers above a river, reflecting on the water with beautiful lighting.
Row 1, Col 5	A castle with towers and catapults is heading towards an iceberg at dawn in a highly detailed painting.
Row 2, Col 1	A Chinese painting of Shanghai featuring blooming sakura trees.
Row 2, Col 2	One slice of simple cheese pizza on a paper plate.
Row 2, Col 3	A throw rug on the floor.
Row 2, Col 4	Graffiti of a rocket ship on a brick wall.
Row 2, Col 5	A young botanist surrounded by pressed flowers and field notes.
Row 3, Col 1	A detailed portrait of a Border Collie with ominous atmosphere.
Row 3, Col 2	A small house in the wilderness.
Row 3, Col 3	An office break room with a table.
Row 3, Col 4	A living room filled with books furniture and a flat screen TV.
Row 3, Col 5	A wooden deck overlooking a mountain valley.
Row 4, Col 1	A kitchen has a stainless steel refrigerator and other appliances.
Row 4, Col 2	A view of the Orion constellation in the night sky.
Row 4, Col 3	Snow mountain and tree reffection in the lake.
Row 4, Col 4	An old oil painting of Dubrovnik on canvas.
Row 4, Col 5	A white country home with a wrap-around porch.
Row 5, Col 1	Sunset over the sea.
Row 5, Col 2	Teacups surounding a kettle.
Row 5, Col 3	A diver swimming beside an enormous manta ray.
Row 5, Col 4	A rabbit sitting on a turtle's back.
Row 5, Col 5	A young badger delicately snifffng yellow roses, richly textured oil painting.

Table 6: Detailed prompts used for generated images in Figure 9.

Image	Prompt
Row 1, Col 1	An eagle.
Row 1, Col 2	A room with two chairs and paintings.
Row 1, Col 3	Splashing milk and berries in a porcelain bowl, vintage wooden tabletop, black background, soft light.
Row 1, Col 4	A moonlit desert canyon beneath a sky full of stars.
Row 1, Col 5	Mist rising over a pine forest at sunrise.
Row 2, Col 1	Lavender fields stretching toward distant hills.
Row 2, Col 2	A coastal cliff with waves crashing below.
Row 2, Col 3	Northern lights above a frozen lake.
Row 2, Col 4	A quiet meadow filled with red poppies.
Row 2, Col 5	A winding river through a golden autumn valley.
Row 3, Col 1	A volcanic mountain glowing at dusk.
Row 3, Col 2	A bamboo grove after a gentle rain.
Row 3, Col 3	Rolling sand dunes under dramatic evening light.
Row 3, Col 4	A tropical lagoon with clear shallow water.
Row 3, Col 5	A lone lighthouse on a rocky island at sunrise.
Row 4, Col 1	Sunbeams passing through a dense jungle canopy.
Row 4, Col 2	A rain-soaked countryside road lined with cypress trees.
Row 4, Col 3	A field of sunflowers under a bright summer sky.
Row 4, Col 4	A moss-covered stone bridge deep in an ancient forest.
Row 4, Col 5	Pink clouds reflected in a silent mountain lake.
Row 5, Col 1	A hidden beach framed by towering limestone cliffs.
Row 5, Col 2	Cherry blossoms drifting beside a quiet river.
Row 5, Col 3	A colorful hillside town overlooking the Mediterranean Sea.
Row 5, Col 4	A minimalist concrete museum surrounded by reflecting pools.
Row 5, Col 5	A quiet suburban street covered in autumn leaves.

Table 7: Detailed prompts used for generated images in Figure 10.

Image	Prompt
Row 1, Col 1	A floating city built above a vast ocean.
Row 1, Col 2	A medieval village beneath a hilltop castle.
Row 1, Col 3	A futuristic laboratory with glowing glass instruments.
Row 1, Col 4	A peaceful spa interior carved from pale stone.
Row 1, Col 5	A Scandinavian bedroom in soft neutral colors.
Row 2, Col 1	A music room with a grand piano and tall arched windows.
Row 2, Col 2	A dramatic side-profile portrait lit by blue and amber light.
Row 2, Col 3	A child wearing oversized aviator goggles and a leather cap.
Row 2, Col 4	A surfer carrying a board along the beach at dawn.
Row 2, Col 5	Three-quarters front view of a blue 1977 Ford F-150 coming around a curve in a mountain road and looking over a green valley on a cloudy day.
Row 3, Col 1	A red fox standing alert in a snowy meadow.
Row 3, Col 2	A playful otter floating among river reeds.
Row 3, Col 3	A family of capybaras relaxing beside a warm spring.
Row 3, Col 4	A vintage motorcycle parked outside a desert gas station.
Row 3, Col 5	A high-speed train arriving at a modern glass station.
Row 4, Col 1	A classic convertible cruising along a coastal highway.
Row 4, Col 2	A racing car speeding through a wet night circuit.
Row 4, Col 3	An enchanted ship sailing through a sky of golden clouds.
Row 4, Col 4	A cyberpunk courier racing across neon rooftops at midnight.
Row 4, Col 5	A holographic marketplace beneath towering digital billboards.
Row 5, Col 1	An underwater research city illuminated by blue lights.
Row 5, Col 2	A crystalline alien forest under a violet sky.
Row 5, Col 3	A cybernetic samurai standing beneath digital cherry blossoms.
Row 5, Col 4	A samurai crossing a wooden bridge in autumn rain.
Row 5, Col 5	An imperial Chinese garden during a lantern festival.

Table 8: Detailed prompts used for generated images in Figure 11.

Image	Prompt
Row 1, Col 1	A quiet seaside town seen from a grassy hill.
Row 1, Col 2	A celestial train station at the edge of the galaxy.
Row 1, Col 3	A music festival beside the ocean in vibrant anime style.
Row 1, Col 4	A traditional Chinese ink painting of misty mountain peaks.
Row 1, Col 5	A richly textured oil painting of a crowded flower market.
Row 2, Col 1	A street racer standing beside a car under neon lights.
Row 2, Col 2	A luxury wristwatch resting on dark volcanic stone.
Row 2, Col 3	A premium camera lens covered in fine raindrops.
Row 2, Col 4	A silver compass resting on an antique map.
Row 2, Col 5	A kayaker passing beneath towering ice arches.
Row 3, Col 1	Handmade pasta dusted with flour in an Italian kitchen.
Row 3, Col 2	A luxury yacht anchored in a crystal-clear bay.
Row 3, Col 3	A group of students watching meteor showers from a rooftop.
Row 3, Col 4	A luminous oil painting of orchards at sunset.
Row 3, Col 5	A dreamy pastel landscape with pink hills and blue trees.
Row 4, Col 1	A horseback rider crossing a desert during a sandstorm.
Row 4, Col 2	Iridescent soap bubbles floating against a black background.
Row 4, Col 3	A pair of handmade glasses on an architect's blueprint.
Row 4, Col 4	An endless staircase spiraling through a cloudy sky.
Row 4, Col 5	A waterfall flowing upward into a floating mountain.
Row 5, Col 1	A flower-filled bicycle leaning beside a pastel spring doorway.
Row 5, Col 2	A research vessel beneath a sky filled with meteor trails.
Row 5, Col 3	A smiling beautiful sorceress wearing a modest high necked blue suit surrounded by swirling rainbow aurora, hyper-realistic, cinematic, post-production.
Row 5, Col 4	A rover following tracks across a frozen moon.
Row 5, Col 5	An archaeologist uncovering a temple beneath desert sand.



Figure 8: Sample images generated by PrefGeo based on SD3.5-M. Prompt are provided in Table 5



Figure 9: Sample images generated by PrefGeo based on SD3.5-M. Prompt are provided in Table 6



Figure 10: Sample images generated by PrefGeo based on FLUX.1-Dev. Prompt are provided in Table 7



Figure 11: Sample images generated by PrefGeo based on FLUX.1-Dev. Prompt are provided in Table 8