DIRECTIONAL TEXTUAL INVERSION FOR PERSONALIZED TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Textual Inversion (TI) is efficient for text-to-image personalization but often fails on complex prompts. We identify embedding norm inflation as a key cause and show that token semantics are primarily encoded by embedding direction. We propose *Directional Textual Inversion* (DTI), which fixes the embedding magnitude to an in-distribution scale and optimizes only direction with a simple MAP objective using a von Mises-Fisher prior. DTI improves prompt fidelity over existing embedding optimization baselines while maintaining competitive subject similarity. Furthermore, we demonstrate its ease of integration and creative applications.

1 INTRODUCTION

Personalization in text-to-image generation adapts models to user-specific concepts. While parameter fine-tuning methods like DreamBooth (Ruiz et al., 2023) are effective, they are computationally intensive. In contrast, embedding optimization methods, such as Textual Inversion (TI) (Gal et al., 2023a), offer a lightweight alternative by updating only token embeddings. This efficiency has made TI a foundational component in numerous personalization frameworks (Hao et al., 2023; Kumari et al., 2023; Tewel et al., 2023b; Lee et al., 2024).

Despite its utility, TI faces critical limitations. Constraining complex visual concepts to a single vector often compromises prompt fidelity and necessitates lengthy optimization. Recent attempts to enrich the embedding space (Voynov et al., 2023; Alaluf et al., 2023) introduce computational overhead without addressing the underlying optimization dynamics that govern semantic alignment.

This paper presents a systematic analysis of TI, revealing that semantic information is predominantly encoded in the embedding direction. We demonstrate, both theoretically and empirically, that unbounded embedding magnitudes during optimization are a primary source of instability, significantly impairing image-text alignment.

Building on these insights, we introduce **Directional Textual Inversion (DTI)**. Unlike standard TI, DTI decouples magnitude from direction; it maintains the magnitude consistent with pre-trained distributions while focusing optimization exclusively on direction. We formulate this as a Maximum a Posteriori (MAP) estimation problem, utilizing a von Mises-Fisher (vMF) prior to regularize the hyperspherical latent space. This approach preserves the efficiency of TI while ensuring robust semantic alignment.

Our evaluation demonstrates that DTI consistently outperforms conventional TI and enhancements like CrossInit (Pang et al., 2024a), achieving superior semantic fidelity with greater efficiency. Furthermore, our directionally optimized embeddings enable novel applications, such as smooth interpolation between personalized concepts.

2 ANALYZING TOKEN EMBEDDING GEOMETRY

2.1 EMPIRICAL MOTIVATION: DIRECTION ENCODES SEMANTICS

Aligning with previous studies (Mikolov et al., 2013; Pennington et al., 2014), we find that token embeddings encode meaning primarily in direction, not magnitude. We validate this by comparing nearest neighbors: cosine similarity (direction-sensitive) yields semantically coherent neighbors for ‘apple’ (*top-5: apples, fruit, peach, pear, egg*), whereas Euclidean distance (magnitude-sensitive)

(a) Norm inflation (b) Semantic drift

Figure 1: **Empirical Motivation.** Standard TI degrades prompt fidelity due to (a) **excessive norms** compared to the original vocabulary and (b) **semantic drift** away from related concepts. DTI addresses these by preserving both norm scale and directional integrity.

finds unrelated tokens (*top-5:* `U+2069, altrin, lestwe, heartnews, samanthaprabhu`). Figure 1b shows that standard TI often neglects this, causing semantic drift where learned embeddings (e.g., `<cat>`) move away from related concepts. This motivates our directional optimization approach.

2.2 Why large magnitudes lead to low text fidelity

Effect I: Positional information is attenuated (see Lemma 1). After LayerNorm/RMSNorm layer, the normalized signal that feeds attention/MLP becomes less sensitive to small additive terms as m grows. Positional information contributes $\mathcal{O}(1/m)$ to the normalized signal $\mathrm{Norm}(m\boldsymbol{v}+\boldsymbol{p})$. Intuitively, a very large-norm token *forgets where it is in the sequence,* weakening contextualization, resulting in omission of details such as style and background (see Figure 1).

Effect II: Residual updates stagnate (see Lemma 2). The residual updates, $F_\ell(\mathrm{Norm}(\boldsymbol{x}^{(\ell)}))$, are computed from a *normalized* inputs and thus have a bounded magnitude. When this bounded update is added through the skip connection to a large vector $\boldsymbol{x}^{(l)}$, the *relative* change (i.e., turning angle of the hidden state's direction) becomes tiny, decreasing in proportion to $1/\|\boldsymbol{x}^{(l)}\|$. In other words, large-norm hidden states become *stuck* in their direction and are difficult for subsequent layers to refine. This *residual stagnation* accumulates across layers, severely limiting the total directional change the initial token can undergo, as formalized in the following proposition and corollary.

Proposition 1 (Accumulated directional drift across L pre-norm blocks)**.** *Let* $\boldsymbol{x}^{(0)} \neq \mathbf{0}$ *and* $\boldsymbol{x}^{(\ell+1)} = \boldsymbol{x}^{(\ell)} + F_\ell(\mathrm{Norm}(\boldsymbol{x}^{(\ell)}))$ *for* $\ell = 0,\dots,L-1$. *Let* $B_\ell := \sup_{\boldsymbol{u}\in S}\|F_\ell(\boldsymbol{u})\|_2 < \infty$, *and* $S_L := \sum_{j=0}^{L-1} B_j$. *Assume* $\|\boldsymbol{x}^{(0)}\|_2 > S_L$, *then*

$$\angle\big(\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(L)}\big) \;\le\; \frac{\pi}{2}\sum_{\ell=0}^{L-1}\frac{B_\ell}{\|\boldsymbol{x}^{(0)}\|_2 - \sum_{j<\ell} B_j} \;\le\; \frac{\pi}{2}\frac{S_L}{\|\boldsymbol{x}^{(0)}\|_2 - S_L}.$$

Corollary 1 (Scaling $\Rightarrow$ directional freezing)**.** *With the notation of Proposition 1, for any* $\alpha > 1$,

$$\angle(\alpha\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(L)}(\alpha)) \;\le\; \frac{\pi}{2}\frac{S_L}{\alpha\left\|\boldsymbol{x}^{(0)}\right\| - S_L}0 \;\xrightarrow[\alpha\to\infty]{}\; 0,$$

where $\boldsymbol{x}^{(L)}(\alpha)$ *denotes the depth-*L *output when the initial token is* $\alpha\boldsymbol{x}^{(0)}$.

Together, this explains TI's reduced text fidelity: excessive token magnitude inhibits contextual integration, causing the personalized subject to overshadow prompt constraints such as style and background. This necessitates the explicit magnitude control framework introduced in the subsequent section. We provide formal proofs in Appendix C and empirical validations in Appendix D.

3 Method: Directional Textual Inversion

Based on our observation and analysis on previous section that token embeddings exhibit strong directional characteristics, we introduce *Directional Textual Inversion* (DTI), a framework that

optimizes an embedding's direction with in-distribution norm to enhance text fidelity in personalized text-to-image generation.

3.1 Optimizing only direction on the hypersphere

We reformulate TI by decoupling the magnitude and direction of the learnable token embedding $\boldsymbol{e} \in \mathbb{R}^d$. The embedding can be expressed as:

$$\boldsymbol{e} \;=\; m^\star \boldsymbol{v}, \qquad \boldsymbol{v} \in \mathbb{S}^{d-1}. \tag{1}$$

Here, $\mathbb{S}^{d-1} = \{\boldsymbol{u} \in R^d : \|\boldsymbol{u}\|_2 = 1\}$ denotes the unit sphere. We fix the magnitude $m^\star$ and optimize only the direction ($\boldsymbol{v}$). Specifically, we set $m^\star$ to be an *in-distribution* magnitude derived from the frozen vocabulary of text encoder (e.g., the average norm). In this way, optimization focuses on semantic information in direction while avoiding out-of-distribution (OOD) norms.

Since the parameter space is the unit sphere, Euclidean updates drift off-manifold, making AdamW (Loshchilov & Hutter, 2017) (default optimizer used in TI-like methods) not suitable. To solve this, we use Riemannian stochastic gradient descent (RSGD) (Bonnabel, 2013) with tangent-space projection and retraction. See Algorithm 1 and Appendix E.1 for further details.

3.2 Maximum A Posteriori formulation with a directional vMF prior

To incorporate directional prior, we formulate the optimization for the optimal direction $\boldsymbol{v}^*$ as a Maximum A Posteriori (MAP) estimation problem. Given a dataset of images $\mathcal{D} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$, the MAP estimate is found by maximizing the posterior probability:

$$\boldsymbol{v}^* = \arg\max_{\mathbf{v}} p(\mathbf{v} \mid \mathcal{D}) \propto \arg\max_{\mathbf{v}} \left[\log p(\mathcal{D} \mid \mathbf{v}) + \log p(\mathbf{v})\right]. \tag{2}$$

Minimizing the negative log-posterior is equivalent to minimizing a loss function composed of a data term and a prior term: $\mathcal{L}(\boldsymbol{v}) = \mathcal{L}_{\text{data}}(\boldsymbol{v}) + \mathcal{L}_{\text{prior}}(\boldsymbol{v})$.

The data term, $\mathcal{L}_{\text{data}} = -\log p(\mathcal{D} \mid \mathbf{v})$, is the negative log-likelihood of the images given the direction. Following standard practice for diffusion models (Ho et al., 2020), we use the mean squared error (MSE) between the true and predicted noise as the objective: $\mathcal{L}_{\text{data}}(\boldsymbol{v}) \coloneqq \mathbb{E}_{\boldsymbol{z},t,\boldsymbol{\epsilon},c}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, c(\boldsymbol{v}))\|_2^2]$. Here, $\boldsymbol{\epsilon}_\theta$ and $c(\cdot)$ are the diffusion model and text encoder, respectively. The Euclidean gradient of this objective, $\boldsymbol{g}_{\text{euc}} = \nabla_{\boldsymbol{v}}\mathcal{L}$, is used in the RSGD update.

For the prior term, $-\log p(\mathbf{v})$, we use a von Mises-Fisher (vMF) distribution on the direction $\boldsymbol{v}$ (detailed justification in Appendix E.2). The vMF distribution is a probability distribution on the $(d-1)$-sphere, analogous to the Gaussian distribution in Euclidean space. It is parameterized by a mean direction $\boldsymbol{\mu} \in \mathcal{S}^{d-1}$ and a concentration parameter $\kappa \geq 0$. The probability density function is given by:

$$p(\mathbf{v}|\boldsymbol{\mu}, \kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \exp(\kappa \boldsymbol{\mu}^\mathsf{T} \boldsymbol{v}), \tag{3}$$

where $I_{d/2-1}$ is the modified Bessel function of the first kind. Here, we work with unnormalized density: $p(\mathbf{v}) \propto \exp(\kappa \boldsymbol{\mu}^\mathsf{T} \boldsymbol{v})$. Ignoring constants, the negative log-prior yields our regularization term, $\mathcal{L}_{\text{prior}}(\boldsymbol{v}) = -\kappa \boldsymbol{\mu}^\mathsf{T} \boldsymbol{v}$.

Selection of vMF parameters. We set mean direction $\boldsymbol{\mu}$ to the constant, normalized embedding of the corresponding class token (e.g., 'dog'). Since estimating concentration parameter κ is non-trivial, we treat it as a hyperparameter that controls the strength of the prior. Based on a grid search (optimal range `5e-5` to `2e-4`), we fixed κ to `1e-4` for all experiments. See Appendix F.2 and F.4 for details.

4 Experiments

4.1 Experimental setups

We evaluated our method using all reference images from the DreamBooth dataset (Ruiz et al., 2023) across 40 prompts, including 10 complex additions. Our primary model was Stable Diffusion XL (SDXL) (Podell et al., 2024), with additional experiments on SANA 1.5 (Xie et al., 2025) to demonstrate applicability to recent architectures. We compared DTI primarily against Textual Inversion (TI) (Gal et al., 2023a) and CrossInit (Pang et al., 2024a), deferring other baselines to Appendix F.3. Following established protocols (Ruiz et al., 2023; Kumari et al., 2023), we measured subject fidelity using DINOv2 (Oquab et al., 2023) cosine similarity and image-text alignment using SigLIP (Zhai et al., 2023). Implementation details are in Appendix F.1.

Table 1: Our DTI consistently improves baselines by generating outputs with enhanced text fidelity while maintaining subject similarity.

Methods	SDXL		SANA 1.5-1.6B		SANA 1.5-4.8B	
	Image	Text	Image	Text	Image	Text
TI	**0.561**	0.292	**0.480**	0.621	0.446	0.646
TI-rescaled	0.243	0.466	0.253	0.655	0.287	0.548
CrossInit	0.545	0.464	0.344	0.614	0.299	0.622
DTI (ours)	0.450	**0.522**	0.479	**0.744**	**0.452**	**0.757**

Table 2: Ablation studies. We tested and confirmed the effectiveness of every component of our DTI.

Optimizer	$m^\star$	$\kappa \times 10^{-3}$	Image	Text
AdamW	mean	0.1	0.335	0.463
RSGD	min	0.1	0.030	0.074
RSGD	5.0 (OOD)	0.1	0.383	0.373
RSGD	mean	0.0	**0.507**	0.436
RSGD	mean	0.5	0.278	**0.688**
RSGD	mean	0.1	0.450	0.522

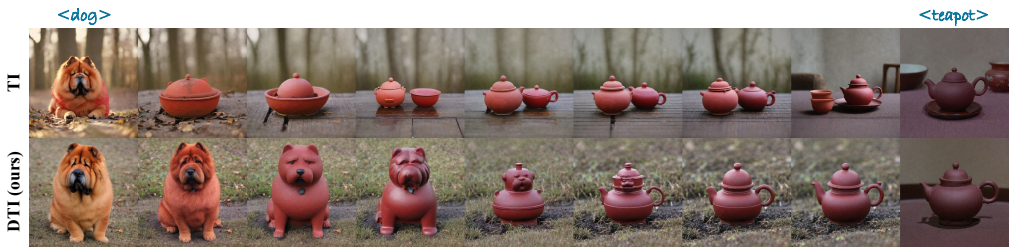


Figure 2: **Subject interpolation.** DTI enables smooth, coherent conceptual transitions, expanding creative personalization capabilities.

4.2 Main Results

We present the main results in the main paper, while more experiments including LoRA fine-tuning, stylization, my object in my style, and face personalization can be found in Appendix F.

Quantitative results. In Table 1, we quantitatively evaluate DTI along two axes: subject similarity and text–prompt fidelity. Overall, the results clearly demonstrates the advantage of DTI over the baselines. Additional comparisons with further baselines on other Stable Diffusion variants are provided in Appendix F.3.

Ablation study. We performed ablation study to verify the effectiveness of components of our DTI, including the optimization space, the embedding magnitude m, and the concentration parameter of vMF distribution κ. The results are summarized in Table 2, with further analyses in Appendix F.4.

Human evaluation. To further examine the effectiveness of our method, we conducted a large scale user study (100 participants via *Amazon Mechanical Turk*) to measure real-world user preferences, with the details in Appendix F.5.

Embedding interpolation for creative applications. We demonstrate DTI's creative potential through embedding interpolation experiments (Figure 2). Leveraging its unit-spherical embedding space, DTI enables coherent transitions via spherical linear interpolation (SLERP), unlike the linear interpolation used by TI which often fails. DTI seamlessly blends distinct concepts, such as merging a dog and a teapot into a hybrid object, or smoothly transitioning between animals like a dog and a cat. It also excels in nuanced face personalization, plausibly interpolating between a young boy and an older woman while maintaining coherence. These results highlight DTI as a powerful tool for intuitive concept blending and creative applications. Additional results are in Appendix F.7.

5 Discussion & Conclusion

As DTI focuses on text prompt fidelity, for applications requiring higher subject fidelity, we recommend integrating DTI with lightweight fine-tuning methods. Empirically, we show our DTI effectively complements lightweight fine-tuning methods like LoRA in Figure 5 and 10.

In this paper, we identified embedding norm inflation as a critical bottleneck for text-to-image alignment. By leveraging the underexplored directional characteristics of the token embedding space, we introduced DTI which optimizes embedding direction while constraining the norm to an in-distribution scale via a novel directional prior. DTI demonstrably enhances prompt fidelity, offering a robust solution for precise and controllable personalization in generative AI.

REFERENCES

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, September 2013. ISSN 0018-9286, 1558-2523.

Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *International Conference on Learning Representations*, 2023a.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *Advances in Neural Information Processing Systems*, 2023b.

Minhyung Cho and Jaehyung Lee. Riemannian approach to batch normalization. In *Advances in Neural Information Processing Systems*, October 2017.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

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 *International Conference on Learning Representations*, 2023a.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023b.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7323–7334, 2023.

Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. ViCo: Plug-and-play visual condition for personalized text-to-image generation, 2023.

Geoffrey E Hinton, Nitish Srivastava, and Kevin Swersky. Lecture 6e: Rmsprop. Coursera, Nonlinear aural component analysis of the cochlea, 2012. URL http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020.

Shoaib Jameel and Steven Schockaert. Word and document embedding with vmf-mixture priors on context word vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023.

Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct Consistency Optimization for Compositional Text-to-Image Personalization. In *Advances in Neural Information Processing Systems*, 2024.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, 2023.

Xianming Li and Jing Li. Aoe: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1825–1839, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. Spherical text embedding. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, July 2023. ISSN 2835-8856.

Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, and Xudong Mao. Cross initialization for face personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. In *Advances in Neural Information Processing Systems*, 2024b.

NaHyeon Park, Kunhee Kim, and Hyunjung Shim. TextBoost: Towards One-Shot Personalization of Text-to-Image Models via Fine-tuning Text Encoder. *arXiv preprint arXiv:2409.08248*, 2024.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Nethods in Natural Language Processing*, pp. 1532–1543, 2014.

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. In *International Conference on Learning Representations*, 2024.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *Advances in Neural Information Processing Systems*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *Proceedings of the International Conference on Machine Learning*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6527–6536, 2024.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023a.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-Locked Rank One Editing for Text-to-Image Personalization. In *Proc. SIGGRAPH*, SIGGRAPH '23, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 979-8-4007-0159-7.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended Textual Conditioning in Text-to-Image Generation, March 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.

Feize Wu, Yun Pang, Junyi Zhang, Lianyu Pang, Jian Yin, Baoquan Zhao, Qing Li, and Xudong Mao. Core: Context-regularized text embedding learning for text-to-image personalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025.

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.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research*, August 2022. ISSN 2835-8856.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems*, 2019.

Dingyi Zhang, Yingming Li, and Zhongfei Zhang. Deep metric learning with spherical embedding. In *Advances in Neural Information Processing Systems*, 2020.

Xulu Zhang, Xiao-Yong Wei, Jinlin Wu, Tianyi Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Compositional inversion for stable diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a.

Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4764–4774, 2024b.

A Related Work

A.1 Personalized text-to-image generation

Recent advancements in text-to-image (T2I) generation have considerably expanded the creative capabilities and flexibility of generative models (Ramesh et al., 2021; Rombach et al., 2022; Nichol et al., 2022; Ramesh et al., 2022; Yu et al., 2022; Podell et al., 2024). Despite these innovations, natural language inherently struggles to precisely convey nuanced, user-specific concepts. This inherent limitation has driven the development of personalization methods, which allow users to generate images reflecting unique concepts with creative prompts.

Textual Inversion (Gal et al., 2023a), which is most well-known for its lightweight integration to many other personalization works, uses embedding optimization by introducing learnable tokens for personalized information without model modification. Subsequent work explored diverse embedding strategies (Voynov et al., 2023; Alaluf et al., 2023; Wu et al., 2025; Zhang et al., 2024a), often with demanding excessive computational costs. Among them, CrossInit (Pang et al., 2024a) offered an efficient initialization strategy with minimal overhead, replacing initialization tokens with the output of text encoder and using regularization loss.

In contrast, fine-tuning based methods such as DreamBooth (Ruiz et al., 2023) achieve high subject fidelity, but require significant computational resources compared to embedding optimization methods (Kumari et al., 2023; Han et al., 2023; Gu et al., 2023; Chen et al., 2023a; Tewel et al., 2023a; Zhang et al., 2024b; Qiu et al., 2023; Pang et al., 2024b) . More recently, Park et al. (2024) proposed fine-tuning text encoder instead of image generator for efficiency, but they still demand more parameters compared to embedding optimization methods.

Meanwhile, there exists a line of encoder-based approaches (Wei et al., 2023; Ruiz et al., 2024; Ye et al., 2023; Gal et al., 2023b; Chen et al., 2023b; Li et al., 2023; Pang et al., 2024b; Ma et al., 2024) that offer fast inference, but they necessitate substantial pre-training.

A.2 Directional embedding space

A number of prior works has emphasized constraining embedding representations to the hypersphere. These include using vMF mixtures for directional clustering (Jameel & Schockaert, 2019), normalizing norms for face recognition (Meng et al., 2019), angle-optimized embeddings to address cosine saturation (Li & Li, 2024), and spherical constraints for uniform document clustering (Zhang et al., 2020). Wang & Isola (2020) offered theoretical support for hyperspherical constraints in contrastive learning. Our method aligns with this trend by modeling embeddings as directional distributions but uniquely decomposes and explicitly optimizes textual embedding direction using a vMF prior within Textual Inversion framework.

B Embedding norm and direction

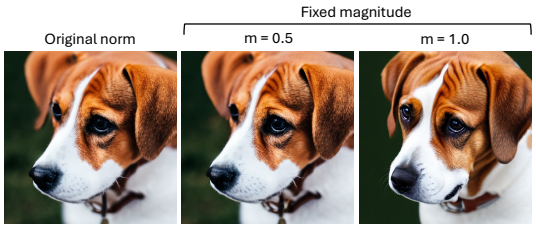


Figure 3: **Effect of magnitude change.** We set the magnitude to a fixed value to analyze the impact of magnitude changes. The resulting outputs show no noticeable difference.

We altered the magnitude of the token as exemplified in Figure 3. However, the resulting output remained mostly unchanged. This indicates that minor adjustments to the magnitude do not significantly affect the outcome.

Table 3: **Nearest tokens under different measures.** We show the nearest tokens to the query words 'study' and 'writing' using both cosine similarity and Euclidean distance.

Query	Cosine	Euclidean
study	studies, studying, research, bookclub, reading, studied, sketches, measurements, thumbnail	U+3160, texanscheer, asober, instaweatherpro, mydayin, premiosmtvmiaw, tairp, thepersonalnetwork, U+2412
writing	writer, write, written, writ, writers, writings, recording, blogging, wrote	phdlife, poetryday, tomorrowspaper, urstrulymahesh, @___, twitterkurds, asober, fakespeare, jamiedor

In Table 3, we provide additional examples illustrating the nearest words retrieved for each query under different similarity measures, which strongly correlate with either direction or magnitude. Our analysis reveals that cosine similarity retrieves words that share semantic meaning with the query. Conversely, Euclidean distance is significantly affected by embedding magnitude, often retrieving words with limited or no semantic relevance. This demonstrates that semantic meaning is predominantly associated with embedding direction rather than magnitude. Note that words beginning with U+ denote Unicode.

C Proofs for theoretical statements

C.1 Setup

Pre-norm block. We study *pre-norm* Transformer blocks

$$\boldsymbol{x}^{(\ell+1)} = \boldsymbol{x}^{(\ell)} + F_\ell(\mathrm{Norm}(\boldsymbol{x}^{(\ell)})), \quad \ell = 0, \dots, L-1, \tag{4}$$

where $\mathrm{Norm} \in \{\mathrm{LayerNorm}, \mathrm{RMSNorm}\}$ (with optional affine (γ, β) absorbed into F_ℓ).

Scale invariance. For normalizations, we use the standard, scale-invariant definitions:

$$\mathrm{RMSN}(\boldsymbol{x}) = \sqrt{d}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}, \quad \mathrm{LN}(\boldsymbol{x}) = \sqrt{d}\frac{\boldsymbol{C}\boldsymbol{x}}{\|\boldsymbol{C}\boldsymbol{x}\|_2}, \quad \boldsymbol{C} := \boldsymbol{I} - \frac{1}{d}\mathbf{1}\mathbf{1}^\top. \tag{5}$$

Thus $\mathrm{RMSN}(s\boldsymbol{x}) = \mathrm{RMSN}(\boldsymbol{x})$ and $\mathrm{LN}(s\boldsymbol{x}) = \mathrm{LN}(\boldsymbol{x})$ for all $s > 0$. Please refer to original papers (Ba et al., 2016; Zhang & Sennrich, 2019) for further details.

Token decomposition. For the input token, we denote $\boldsymbol{x}^{(0)} = m\boldsymbol{v} + \boldsymbol{p}$ with $m > 0$, $\|\boldsymbol{v}\|_2 = 1$, and (optional) absolute positional embedding $\boldsymbol{p} \in \mathbb{R}^d$.

Bounded sub-layers. Define $\mathcal{S} = \{\mathrm{Norm}(\boldsymbol{z}) : \boldsymbol{z} \neq \mathbf{0}\}$. Since Norm maps into a fixed scale, bounded set and F_ℓ (attention/MLP plus projections) is continuous on bounded sets,

$$B_\ell := \sup_{\boldsymbol{u} \in \mathcal{S}} \|F_\ell(\boldsymbol{u})\|_2 < \infty. \tag{6}$$

Throughout, $\|\cdot\|_2$ denotes the Euclidean (l_2) norm.

C.2 Positional attenuation

Lemma 1 (Absolute positional embedding attenuates as $m \to \infty$)**.** *Let $\boldsymbol{x}^{(0)} = m\boldsymbol{v} + \boldsymbol{p}$ with $\|\boldsymbol{v}\|_2 = 1$, $m > 0$, and absolute positional embedding $\boldsymbol{p} \in \mathbb{R}^d$. Suppose* $\mathrm{Norm} \in \{$*LayerNorm, RMSNorm*$\}$ *and $\boldsymbol{v}$ is non-degenerate for LayerNorm (i.e., its per-feature variance is nonzero; this holds for generic token embeddings). Then*

$$\big\|\mathrm{Norm}(m\boldsymbol{v} + \boldsymbol{p}) - \mathrm{Norm}(m\boldsymbol{v})\big\|_2 = \mathcal{O}\Big(\frac{\|\boldsymbol{p}\|_2}{m}\Big) \quad \text{as } m \to \infty \text{ (with } \boldsymbol{v}, \boldsymbol{p} \text{ fixed).}$$

Hence the positional contribution shrinks linearly in $1/m$.

Proof. By scale invariance, $\mathrm{Norm}(m\boldsymbol{v} + \boldsymbol{p}) = \mathrm{Norm}\big(\boldsymbol{v} + \varepsilon\big)$ with $\varepsilon := \boldsymbol{p}/m$, and $\mathrm{Norm}(m\boldsymbol{v}) = \mathrm{Norm}(\boldsymbol{v})$.

RMSNorm. With $\|\boldsymbol{v}\| = 1$,

$$\frac{\boldsymbol{v} + \varepsilon}{\|\boldsymbol{v} + \varepsilon\|} = \boldsymbol{v} + (\boldsymbol{I}_d - \boldsymbol{v}\boldsymbol{v}^\top)\varepsilon + \mathcal{O}(\|\varepsilon\|^2),$$

hence $\mathrm{RMSN}(\boldsymbol{v}+\varepsilon) - \mathrm{RMSN}(\boldsymbol{v}) = \sqrt{d}\,(\boldsymbol{I}_d - \boldsymbol{v}\boldsymbol{v}^\top)\varepsilon + \mathcal{O}(\|\varepsilon\|^2)$ and $\|\mathrm{RMSN}(m\boldsymbol{v}+\boldsymbol{p}) - \mathrm{RMSN}(m\boldsymbol{v})\| \le \sqrt{d}\,\|\boldsymbol{p}\|/m + O(m^{-2})$.

LayerNorm. Write $\boldsymbol{a} := \boldsymbol{C}\boldsymbol{v} \neq \boldsymbol{0}$, $\boldsymbol{u} := \boldsymbol{a}/\|\boldsymbol{a}\|$. Then

$$\frac{\boldsymbol{a}+\boldsymbol{C}\varepsilon}{\|\boldsymbol{a}+\boldsymbol{C}\varepsilon\|} = \boldsymbol{u} + \frac{(\boldsymbol{I}_d - \boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{C}\varepsilon}{\|\boldsymbol{a}\|} + \mathcal{O}(\|\varepsilon\|^2),$$

so $\|\mathrm{LN}(m\boldsymbol{v}+\boldsymbol{p}) - \mathrm{LN}(m\boldsymbol{v})\| = \sqrt{d}\,\frac{(\boldsymbol{I}_d - \boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{C}\boldsymbol{p}}{m\|\boldsymbol{C}\boldsymbol{v}\|} + \mathcal{O}(m^{-2})$, which is $\mathcal{O}(\|\boldsymbol{p}\|/m)$. □

C.3 Residual stagnation

Lemma 2 (Residual stagnation in a pre-norm block)**.** *Let* $\boldsymbol{x}^{(\ell+1)} = \boldsymbol{x}^{(\ell)} + F_\ell(\mathrm{Norm}(\boldsymbol{x}^{(\ell)}))$ *with* $\boldsymbol{x}^{(\ell)} \neq \boldsymbol{0}$ *and* $\mathrm{Norm} \in \{\mathrm{LN}, \mathrm{RMSN}\}$*, and let*

$$B_\ell := \sup_{\boldsymbol{u}\in\mathcal{S}} \|F_\ell(\boldsymbol{u})\|_2 < \infty.$$

Then

$$\frac{\|\boldsymbol{x}^{(\ell+1)} - \boldsymbol{x}^{(\ell)}\|_2}{\|\boldsymbol{x}^{(\ell)}\|_2} \le \frac{B_\ell}{\|\boldsymbol{x}^{(\ell)}\|_2}, \qquad \angle(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(\ell+1)}) \le \arcsin\Big(\frac{B_\ell}{\|\boldsymbol{x}^{(\ell)}\|_2}\Big).$$

Proof. Since $\mathrm{Norm}(\boldsymbol{x}^{(\ell)}) \in S$, we have $\|\boldsymbol{x}^{(\ell+1)} - \boldsymbol{x}^{(\ell)}\|_2 = \|F_\ell(\mathrm{Norm}(\boldsymbol{x}^{(\ell)}))\|_2 \le B_\ell$, giving the first bound. Write $\boldsymbol{x}^{(\ell+1)} = \boldsymbol{x}^{(\ell)} + \delta$. The orthogonal component of δ is at most $\|\delta\|$; a short calculation shows $\sin\angle(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(\ell+1)}) \le \|\delta\|_2/\|\boldsymbol{x}^{(\ell)}\|_2 \le B_\ell/\|\boldsymbol{x}^{(\ell)}\|_2$, which implies the stated angle bound. □

Proposition 1 (Accumulated directional drift across L pre-norm blocks)**.** *Let* $\boldsymbol{x}^{(0)} \neq \boldsymbol{0}$ *and* $\boldsymbol{x}^{(\ell+1)} = \boldsymbol{x}^{(\ell)} + F_\ell(\mathrm{Norm}(\boldsymbol{x}^{(\ell)}))$ *for* $\ell = 0, \dots, L-1$*. Let* $B_\ell := \sup_{\boldsymbol{u}\in S} \|F_\ell(\boldsymbol{u})\|_2 < \infty$*, and* $S_L := \sum_{j=0}^{L-1} B_j$*. Assume* $\|\boldsymbol{x}^{(0)}\|_2 > S_L$*, then*

$$\angle\big(\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(L)}\big) \;\le\; \frac{\pi}{2}\sum_{\ell=0}^{L-1} \frac{B_\ell}{\|\boldsymbol{x}^{(0)}\|_2 - \sum_{j<\ell} B_j} \;\le\; \frac{\pi}{2}\frac{S_L}{\|\boldsymbol{x}^{(0)}\|_2 - S_L}.$$

Proof. Let $\theta_\ell := \angle(x^{(\ell)}, x^{(\ell+1)})$. By the recall above, $\theta_\ell \le \arcsin\big(B_\ell/\big\|x^{(\ell)}\big\|\big) \le \frac{\pi}{2}\,B_\ell/\big\|x^{(\ell)}\big\|$. Also $\big\|x^{(\ell)}\big\| \ge \big\|x^{(0)}\big\| - \sum_{j<\ell} B_j$ (each step can shrink the norm by at most B_ℓ). Summing angles (spherical triangle inequality) gives the first display; since $\big\|x^{(0)}\big\| - \sum_{j<\ell} B_j \ge \big\|x^{(0)}\big\| - S_L$, each fraction is $\le B_\ell/(\big\|x^{(0)}\big\| - S_L)$, yielding the last bound. □

D Empirical Validation of Token Geometry Analysis

We empirically validate the two theoretical effects introduced in the previous sections. Effect I describes the attenuation of positional information under large embedding magnitudes, while Effect II concerns residual-update stagnation in pre-norm Transformer blocks. Our experiments directly probe both behaviors on the base encoder, TI, and our proposed DTI.

Effect I (Attenuation of positional information). We evaluate whether increasing embedding magnitude makes positional information unrecoverable after the first pre-norm normalization (LN). To validate this, we train a 2-layer MLP classifier on the *frozen* base text encoder to predict a token's absolute position from the output of $\mathrm{LN}(\mathbf{e}+\mathbf{p})$, where $\mathbf{e}$ and $\mathbf{p}$ each denotes token and positional embeddings. On unmodified inputs ('Normal' in Figure 4), the classifier achieves 100% accuracy, confirming that LN preserves positional information. We then scale the norm of a single token embedding by a factor $m \in \{0.5, 1, 2, 4, 8, 16\}$ before applying LN. Accuracy deteriorates rapidly once m exceeds the natural scale of the encoder. Furthermore, we evaluated the classifier on TI-trained and DTI-trained personalized embeddings. TI embeddings, which has excessively large norms, collapse to near-zero positional accuracy, while DTI embeddings remain fully recoverable.

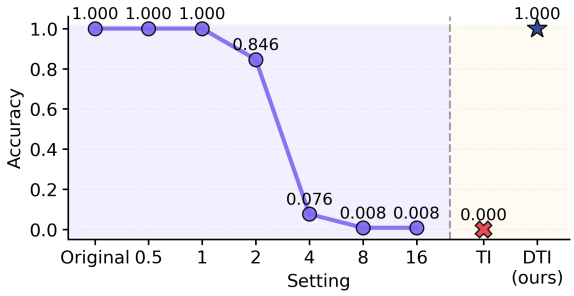


Figure 4: **Emperical validation.** Position prediction accuracy from LN outputs under varying embedding magnitudes, compared with trained TI and DTI embeddings.

This behavior directly corroborates Lemma 1: when m becomes large, $\mathrm{LN}(m\mathbf{v} + \mathbf{p})$ becomes dominated by $m\mathbf{v}$, rendering the positional component $\mathbf{p}$ effectively invisible. DTI avoids this failure mode by constraining magnitudes to remain in-distribution.

Effect II (Residual-update stagnation). To test Lemma 2, we measure the internal angular change of hidden states within each pre-norm Transformer block. For each concept token, we compute the angle between the hidden state entering and exiting each block and then average across all layers. The average per-block angular change of TI embeddings was $21.33°$, whereas the angular change of DTI embeddings was $33.52°$ ($\mathbf{1.57}\times$ larger).

These results support the theoretical prediction that excessively large norms suppress the effective residual direction in pre-norm blocks, causing the forward computation to behave nearly as an identity mapping. By keeping embedding norms within the training distribution, DTI prevents such stagnation and permits substantially larger and more meaningful updates throughout the encoder.

E Extended Methods

E.1 RSGD for token embedding optimization

We employ Riemannian Stochastic Gradient Descent (RSGD) to optimize the directional embedding $\boldsymbol{v}$ while ensuring it remains on the unit hypersphere $\mathbb{S}^{d-1}$. The update step at iteration k involves projecting the Euclidean gradient onto the tangent space $T_{\boldsymbol{v}_k}\mathbb{S}^{d-1}$ and retracting the updated vector back to the manifold:

$$\boldsymbol{g} \;=\; \boldsymbol{g}_{\text{euc}} - (\boldsymbol{v}_k^{\mathsf{T}}\boldsymbol{g}_{\text{euc}})\,\boldsymbol{v}_k \;\in T_{\boldsymbol{v}_k}\mathbb{S}^{d-1}, \quad \boldsymbol{v}_{k+1} \;=\; \mathrm{Retr}_{\boldsymbol{v}_k}(-\eta\boldsymbol{g}) \;=\; \frac{\boldsymbol{v}_k - \eta\boldsymbol{g}}{\|\,\boldsymbol{v}_k - \eta\boldsymbol{g}\,\|_2}. \tag{7}$$

Here, $\boldsymbol{g}_{\text{euc}}$ is a Euclidean space gradient, $\boldsymbol{g} \in T_{\boldsymbol{v}_k}\mathcal{S}^{d-1}$ is a tangent-space gradient, and $\eta > 0$ is a learning rate. In practice, we scaled the gradient $\boldsymbol{g}$ using its own norm similarly. This was inspired by Euclidean space optimizers (Hinton et al., 2012; Kingma & Ba, 2015; Loshchilov & Hutter, 2019), which normalizes the gradient based on moving average of squared gradients.

We observe that gradient magnitudes tend to increase as training progresses, which often leads to instability in the later stages. Although standard learning rate schedules can help mitigate this issue, the gradient dynamics vary considerably across different datasets and training settings, limiting the effectiveness of fixed schedules. To address this, we draw inspiration from adaptive optimization techniques in Euclidean space (Kingma & Ba, 2015; Duchi et al., 2011) and propose a simple yet effective gradient scaling scheme based on gradient norms:

$$\boldsymbol{g}'_k = \boldsymbol{g}_k / \|\boldsymbol{g}_k\|_2, \tag{8}$$

where $\boldsymbol{g}_k$ is the gradient at iteration k. This normalization is equivalent to using an adaptive step size $\eta/\|\boldsymbol{g}_k\|$ in a Euclidean update $\boldsymbol{v}_{k+1} = \boldsymbol{v}_k - (\eta/\|\boldsymbol{g}_k\|_2)\boldsymbol{g}_k$; the update direction is still $\boldsymbol{g}_k$, but the length is fixed to η, preventing very large gradients from causing excessively large parameter updates. Note that a similar technique was previously explored in the context of Riemannian optimization (Cho & Lee, 2017).

E.2 WHY VMF OVER OTHER DISTRIBUTIONS?

We chose the von Mises-Fisher (vMF) distribution as it is ideally suited for modeling the directional characteristics of token embeddings we identified in Section 2. Our central hypothesis is that the token embedding vocabulary can be modeled as a **mixture of vMF distributions**, where each component corresponds to a distinct semantic cluster (e.g., one for animals, another for objects). The vMF distribution is the suitable building block for this model for three key reasons:

- **It's a natural fit.** The vMF is the natural analog to the Gaussian distribution on a hypersphere, making it a principled and standard choice for modeling directional data clusters.
- **It's computationally efficient.** The vMF's mathematical form is exceptionally convenient for optimization. In our MAP formulation, the gradient of the log-prior is a *constant-direction vector* ($-\kappa\mu$), which provides a stable and efficient semantic pull without requiring complex calculations. This simplicity makes it more suitable for high-dimensional embeddings in large-scale models than alternatives like the Kent and Bingham distributions.
- **It's interpretable and controllable.** The parameters are easy to understand. The mean direction $\boldsymbol{\mu}$ serves as a *semantic anchor* to prevent the learned token from drifting away from related concepts, while the concentration κ allows us to control the strength of this regularization.

These factors collectively make the vMF distribution a superior choice for our application, providing the necessary regularization in a way that is both mathematically principled and computationally tractable.

F EXTENDED EXPERIMENTS

Algorithm 1 Directional Textual Inversion (DTI)

Inputs: Model $\boldsymbol{\epsilon}_\theta$, text encoder $c(\cdot)$, init token $\boldsymbol{e}_{\text{init}}$, magnitude m^*, κ, iterations K, learning rate η
$\boldsymbol{v}_0 \leftarrow \boldsymbol{e}_{\text{init}} / \|\boldsymbol{e}_{\text{init}}\|_2$
$\boldsymbol{\mu} \leftarrow \boldsymbol{e}_{\text{init}} / \|\boldsymbol{e}_{\text{init}}\|_2$
for $k = 0$ to $K - 1$ **do**
  Sample minibatch $(\boldsymbol{z}, t, \boldsymbol{\epsilon})$
  $\boldsymbol{g}_{\text{data}} \leftarrow \nabla_{\boldsymbol{v}} \mathcal{L}_{\text{data}}(m^* \boldsymbol{v}_k)$
  $\boldsymbol{g}_{\text{euc}} \leftarrow \boldsymbol{g}_{\text{data}} - \kappa\boldsymbol{\mu}$ (add prior gradient)
  $\boldsymbol{g} \leftarrow \boldsymbol{g}_{\text{euc}} - (\boldsymbol{g}_{\text{euc}}^{\mathsf{T}} \boldsymbol{v}_k)\, \boldsymbol{v}_k$ (tangent projection)
  $\boldsymbol{g}' \leftarrow \boldsymbol{g} / \|\boldsymbol{g}\|_2$ (gradient scaling)
  $\boldsymbol{v}_{k+1} \leftarrow \dfrac{\boldsymbol{v}_k - \eta\, \boldsymbol{g}'}{\|\boldsymbol{v}_k - \eta\, \boldsymbol{g}'\|_2}$ (retraction to $\mathcal{S}^{d-1}$)
end for
return $\boldsymbol{e}^* = m^* \boldsymbol{v}_K$

F.1 IMPLEMENTATION DETAILS

Following the protocol of recent studies, we primarily conducted experiments using Stable Diffusion XL (SDXL). To demonstrate broader applicability to different models, we also conducted experiments with very recent model, SANA 1.5 (Xie et al., 2025), where the results can be found in Table 1.

For a fair comparison, we adopted most of the hyperparameter settings from the Textual Inversion (TI) implementation provided by the HuggingFace `diffusers` library. Specifically, we used a training batch size of 4, and enabled mixed-precision training with the bfloat16 (bf16) format. We set the learning rate commonly-used $5e{-}3$. All experiments were run with a fixed random seed of 42, and the maximum number of training steps was set to 500. For output generation, we used the `DDIMScheduler` with 50 inference steps for SDXL and 20 steps with `FlowMatchEulerDiscreteScheduler` for SANA.

Hyperparemeters. There can be various approaches to selecting the concentration parameter κ. We performed a grid search and found that values in the range of `5e-5` to `2e-4` works well. Therefore, we did not conduct a extensive search for an optimal decimal value. Throughout the experiments, we simply fixed value to `1e-4`, which generalizes well to experiments with different settings. Examples illustrating the effects of different κ settings are provided in Table 2.

Baselines. Throughout this paper, we compare our method with two baseline approaches: Textual Inversion (TI) (Gal et al., 2023a) and CrossInit (Pang et al., 2024a). Since the official CrossInit implementation is based on Stable Diffusion v2.1 with hyperparameters tailored to that version, we reconfigure it to operate on SDXL by aligning its training setup with that of TI. Specifically, we adopt the same hyperparameters as used for TI, and we set the regularization weight for CrossInit to $1e-5$, as specified in the original paper.

F.2 On the choice of prior

For all of our experiments in the main section, we used the initial tokens as prior from the DreamBooth dataset as is. However, we would like to note that since our DTI can leverage the prior, searching for better priors can lead to better results. This demonstrates the effectiveness of the prior.

To test this, we experimented with having a VLM recommend initial tokens. More specifically, we provided reference images to the VLM and asked it to recommend 1-2 words that best describe them. For the experiments, we used Qwen-VL 2.5 (Bai et al., 2025) as the VLM. The results are shown in Table 4.

The results indicate that changing the prior affects performance, although the overall effect is modest. For both TI and our DTI, Qwen-VL initialization tends to increase subject similarity, accompanied by a slight decrease in text fidelity. Practitioners may leverage VLMs or manually craft priors with targeted terms to emphasize desired attributes. Overall, these findings demonstrate the flexibility and effectiveness of leveraging priors.

Table 4: Results with VLM-recommended priors. We compare Qwen-VL recommended initial tokens with DreamBooth initial tokens as priors for DTI.

		SDXL		SANA	
Method	**Initialization**	**Image**	**Text**	**Image**	**Text**
TI	DreamBooth init	0.561	0.292	0.480	0.621
	Qwen-VL init	0.583	0.273	0.501	0.619
DTI (ours)	DreamBooth init	0.450	0.522	0.479	0.744
	Qwen-VL init	0.520	0.391	0.504	0.697

F.3 Comparison with other baselines

We expand our comparative analysis to include additional baselines: P+ (Voynov et al., 2023), NeTI (Alaluf et al., 2023), and CoRe (Wu et al., 2025). We run these experiments mainly on SD1.5 and SD2.1-base as these baseline papers work on those versions. Adhering to the evaluation protocol of the main paper, we measure subject similarity using DINOv2 similarity and prompt fidelity with the CLIP-variant, SigLIP. The results demonstrate that across both architectures, DTI consistently achieves the most favorable balance between these metrics compared to all baselines. Qualitative comparison can be found in Figure 9.

DTI as a drop-in replacement for TI. Although DTI is primarily designed for embedding-only personalization, it also functions effectively as a drop-in replacement within fine-tuning pipelines. Recent work on Direct Consistency Optimization (DCO) Lee et al. (2024) typically performs joint optimization of a concept token using standard Textual Inversion (TI). To assess the impact of substituting this TI component with DTI, we conducted joint training for 250 steps using a LoRA module with rank 4.

As shown in Table 6, the conventional TI-based joint training exhibits limited text–alignment (0.456), whereas replacing TI with DTI substantially improves alignment to 0.635. This quantitative improve-

Table 5: **Results on SD1.5 and SD2.1-base.** We compare the baselines that improve TI on different versions of Stable Diffusion. DTI achieves the best balance between subject similarity and text fidelity compared to other baselines.

Method	SD1.5		SD2.1-base	
	Image	Text	Image	Text
P+ (Voynov et al., 2023)	0.273	**0.719**	0.238	**0.663**
NeTI (Alaluf et al., 2023)	0.408	0.579	0.565	0.517
CoRe (Wu et al., 2025)	0.340	0.661	0.357	0.654
DTI (ours)	**0.418**	0.554	<u>0.469</u>	0.568

ment is further reflected in Figure 5, where our method accurately incorporates textual attributes (e.g., red backpack), while DCO with TI fails to do so.

Table 6: **DCO experiments.** Comparison of DCO with standard Textual Inversion (TI) versus DCO initialized with our DTI. Integrating our method significantly improves text alignment.

Method	Image	Text
DCO	**0.605**	0.456
DCO + DTI (ours)	0.568	**0.635**

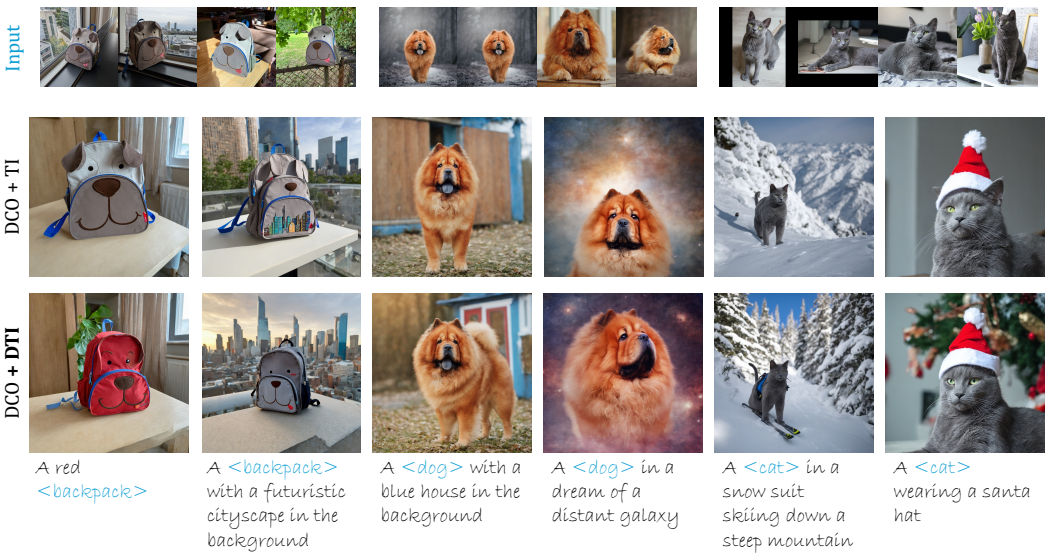


Figure 5: **Qualitative results with DCO (Lee et al., 2024).** We provide the comparison of the output images of DCO + TI and DCO + DTI. The results suggest our DTI's superiority in text fidelity while preserving strong image similarity.

F.4 Ablation Study

Effect of Riemannian optimization. Our DTI framework employs Riemannian optimization to ensure embeddings lie on the spherical manifold $\mathbb{S}^{n-1}$. An alternative is to simply re-scale embeddings after each Euclidean optimization step to achieve this constraint. However, Table 2 (first row) shows this latter Euclidean-based approach with re-scaling achieves suboptimal results, highlighting the benefit of direct Riemannian optimization.

Effect of magnitude (m). We investigated the impact of the fixed embedding magnitude, m, on personalization performance. Our DTI framework, by default, sets m to the average norm observed in the pre-trained CLIP token vocabulary. We compared this "mean" strategy under the Riemannian optimization setting with $\kappa = 1e-4$:

Table 7: We surveyed real-world user preferences regarding subject fidelity and image-text alignment. DTI ranks the top in both metrics, confirming its practical benefits.

	TI	CrossInit	**DTI (ours)**
Image fidelity	13.78	42.87	**43.45**
Text alignment	10.83	22.40	**66.77**

- Setting m to the minimum vocabulary norm ("min").
- Setting m to the mean vocabulary norm ("mean").
- Setting m to a large, out-of-distribution (OOD) value of 5.0.

As shown in Table 2:

- The "mean" strategy achieves the highest subject similarity and strong text fidelity.
- The "min" strategy results in significantly poorer performance in both metrics.
- Using an OOD magnitude of 5.0 also leads to a degradation in both metrics.

These results validate our choice of fixing the magnitude to an in-distribution scale, specifically the average vocabulary norm, as it provides a strong balance of subject similarity and text alignment. Both excessively small ("min") and out-of-distribution large ("OOD") magnitudes are detrimental.

Effect of concentration parameter (κ). The concentration parameter κ of the von Mises-Fisher (vMF) prior controls the strength of the directional regularization. We analyzed its effect by varying κ while using Riemannian optimization and the "mean" embedding magnitude. We tested $\kappa = 0.0$ (no prior), $\kappa = 1e-4$ (DTI default), and $\kappa = 5e-4$.

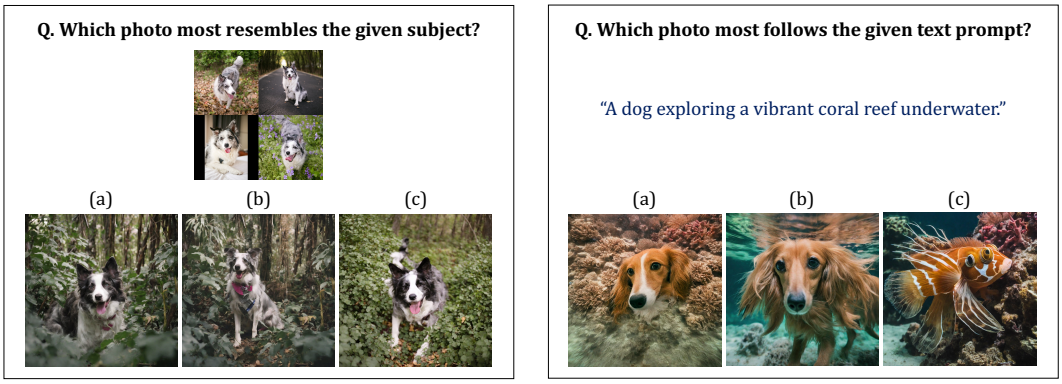
The results in Table 2 indicate:

- With $\kappa = 1e-4$, we observe the best balance between subject similarity and text fidelity.
- Setting $\kappa = 0.0$, which removes the directional prior, leads to lower scores in text fidelity, which validates our method's priority in model's enhancing semantic understanding.
- Increasing the regularization strength with $\kappa = 5e-4$ yields the highest text fidelity among the tested values but at the cost of reduced subject similarity.

Overall, our default choice of $\kappa = 1e-4$ provides a better balance between maintaining subject similarity and ensuring text fidelity. Note that $\kappa = 1e-4$ may not be strictly optimal in decimals across all criteria but works reasonably well by providing robust overall performance.

F.5 User study

To evaluate real-world user preferences for image generation quality, we conducted a comprehensive user study involving 100 participants recruited through Amazon Mechanical Turk. Each participant completed a survey consisting of 20 questions, evenly divided into two critical evaluation criteria: subject similarity and text prompt fidelity. For each question, participants were presented with three distinct image options, generated by: Textual Inversion (Gal et al., 2023a), CrossInit (Pang et al., 2024a), and our proposed Directional Textual Inversion (DTI). The order of these three choices was randomized for each question, using a fixed random seed to ensure consistent shuffling across all participants. Sample questions can be found in Figure 6. We collected a total of 96 valid responses, with 4 submissions being excluded due to invalid patterns such as selecting the same answer for all questions. The results, as detailed in Table 7, demonstrate that our Directional Textual Inversion (DTI) consistently outperforms both Textual Inversion and CrossInit across both evaluation metrics: image subject similarity and text prompt fidelity. These findings confirm the superior performance of our proposed method in generating images that more accurately align with user expectations regarding both visual content and textual descriptions.



(a) Subject similarity (b) Text prompt fidelity

Figure 6: **User study design.** We conducted a user study with 100 participants recruited via Amazon Mechanical Turk to evaluate 20 questions. The evaluation focused on two key aspects: subject similarity (10 questions) and text prompt fidelity (10 questions). To ensure fair comparison, the random seed was fixed and option order was shuffled.

Figure 7: **Qualitative results on SDXL.** We compare DTI with previous methods across diverse subjects and textual prompts, spanning simple descriptions to complex variations in attributes, backgrounds, and styles (same random seeds). All results in this figure are generated with SDXL (SANA in Appendix Figure 8).

F.6 Qualitative results

Figure 7 illustrates qualitative comparisons across various prompts. DTI consistently generates images that more accurately reflect the content of the captions, while effectively preserving subject consistency. For instance, for 'Pop-art style illustration of <cat>', TI omits the cat while DTI renders the cat in the specified style. Similarly, in the second column, TI and CrossInit fail to incorporate all elements of the prompt, disregarding either the subject or details such as 'music stage' and 'spotlight'.

In contrast, DTI integrates both the subject and these details, producing a more complete output. Collectively, these examples highlight DTI's superior compositional fidelity and subject preservation, showing its powerfulness that consistently satisfies all prompt constraints. This attributes to DTI's stable optimization within the directional space, which facilitates improved integration of multiple prompt components. DTI's ability to maintain subject fidelity and adhere to textual intent establishes it as a robust choice for a wide range of text-to-image generation tasks. Additional qualitative results including those of SANA can be found in Appendix F.6.

We present additional qualitative comparisons with TI-based approaches (Gal et al., 2023a; Pang et al., 2024a) in Figure 7 (SDXL) and 8 (SANA). The results illustrate that our proposed DTI consistently generates outputs that accurately align with the provided text prompts, even in challenging cases where the baseline methods fail to do so.

Our DTI serves as a drop-in replacement for TI, enhancing the model's performance when combined with LoRA. The qualitative results in Figure 10 demonstrate that DTI consistently generates outputs that both precisely follow the text prompt and accurately capture the subject's details.

F.7 More results on applications

Stylization. We explore the combination of personalized subject embeddings and style embeddings. Our method, DTI, consistently generates images that accurately reflect both the personalized subject and the specified style. In contrast, TI frequently fails in this task, either by omitting the subject altogether (top row) or by inadequately capturing the intended style or subject details (bottom row) of Figure 11.

My object in my style. We also compare our results in simultaneous generation of personalized subject and style. The results demonstrated in Figure 12 shows that DTI successfully generates outputs that are faithful to both subject and style, while TI fails to.

Face personalization. To evaluate and showcase the capability of our DTI method in face personalization, we conducted experiments using randomly selected faces from the FFHQ dataset (Karras et al., 2019) as well as faces generated by DALL·E (Ramesh et al., 2021).

Since CrossInit specifically focuses on facial personalization, we compare TI, CrossInit and our DTI on this task. Given that CrossInit does not explicitly provide hyperparameters (including learning rate) tailored for SDXL, we performed a grid search across various learning rates. Our empirical results indicated that the learning rate used by TI yielded reasonable performance for CrossInit as well. Figure 14 illustrates a comparison between the three methods, demonstrating that all methods perform effectively for facial personalization. Nevertheless, as the complexity of text prompts increases (rows depicted in the left columns), the baseline methods struggle to accurately reflect all described components of the prompts. In contrast, our DTI method consistently captures the critical components precisely, demonstrating superior performance in achieving enhanced textual fidelity.

More results on interpolation. More results on interpolating the personalized subjects can be found in Fig. 13, demonstrating our creative application capability.

G Additional Experiments

We present additional experimental results in Figures 15, 16, 17, and 18. Specifically, Figure 15 compares TI using SLERP against LERP, justifying our choice of the latter. In Figure 16, we present an ablation study on magnitude settings. While our DTI uses the mean value of the entire vocabulary as the default, we further investigate initializing with the specific category describing the subject (e.g., cat). We demonstrate that minor variations in magnitude do not significantly alter the outcome. Figure 17 evaluates our DTI in multi-concept scenarios, illustrating both successful outcomes and limitations. Finally, we analyze specific failure cases of our method in Figure 18.

H Societal impacts

The rapid advancement of text-to-image diffusion models, especially in the domain of personalization techniques, raises important societal considerations. In particular, the ease of generating highly

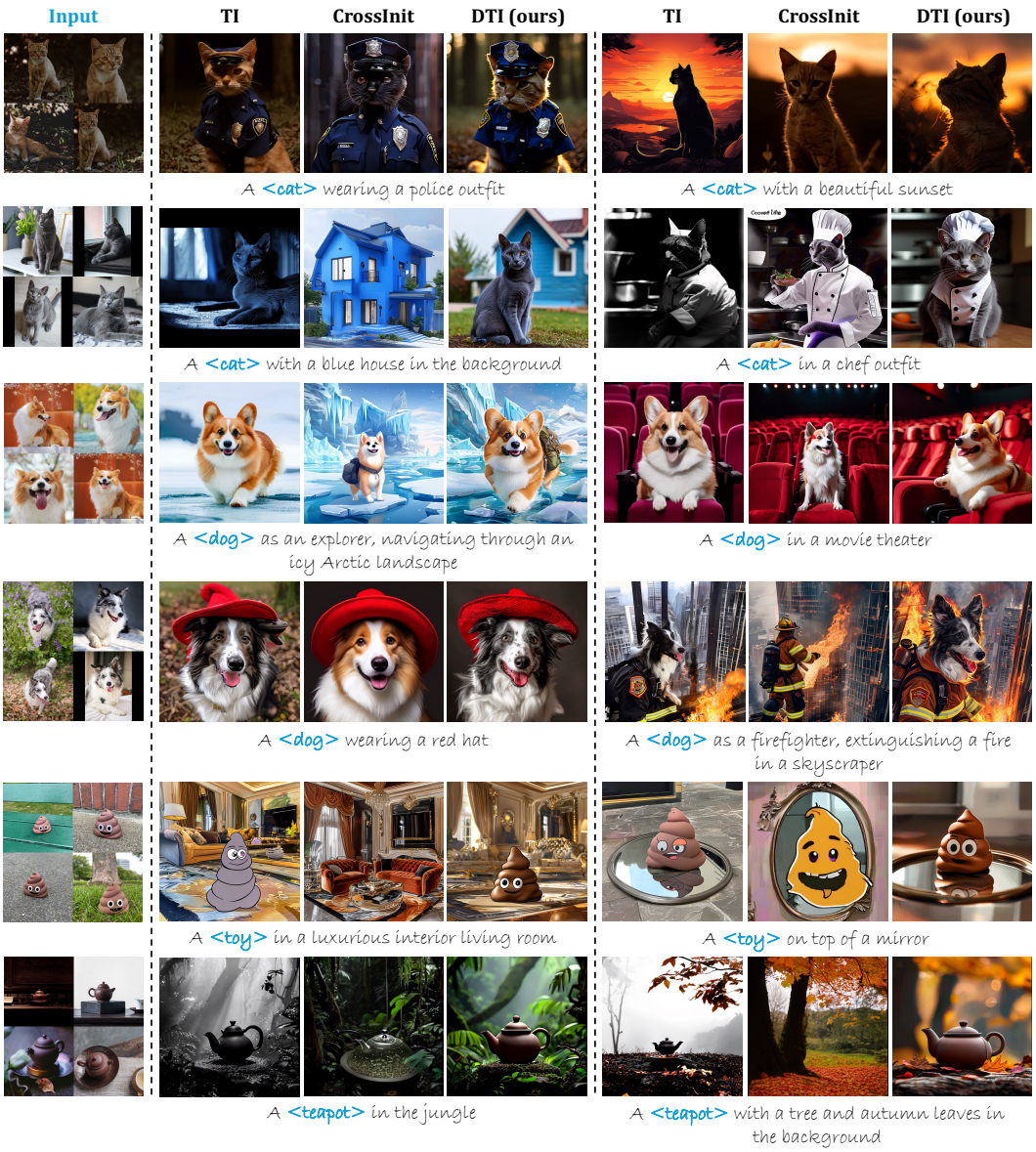


Figure 8: **Qualitative results with SANA1.5-1.6B.** Here, we provide more qualitative comparison with TI and CrossInit on SANA. Our DTI consistently generates results that precisely reflect the user text prompts, maintaining the subject similarity at the same time.

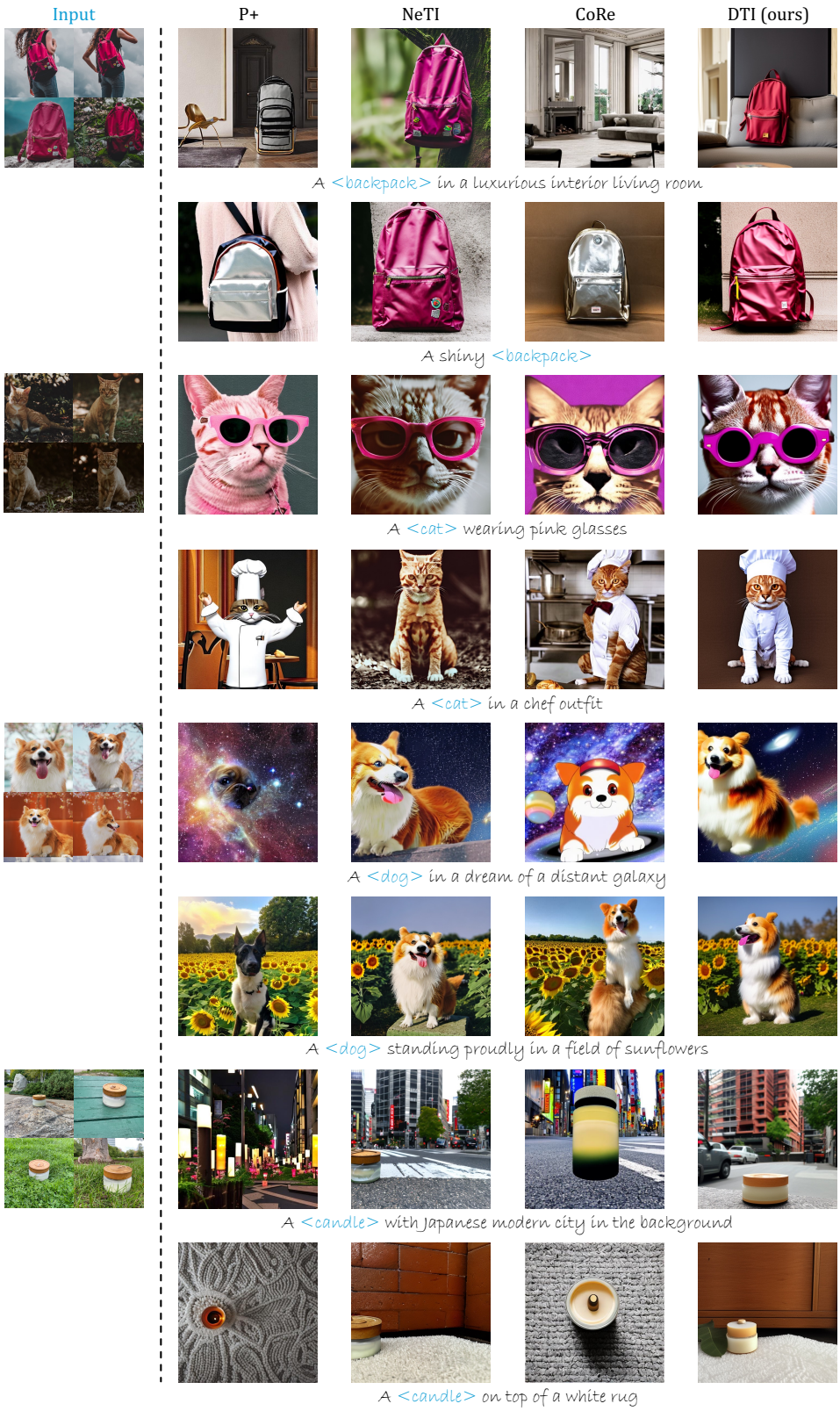


Figure 9: **Comparison with additional baselines.** We provide qualitative comparisons against additional TI-enhancing methods—P+, NeTI, and CoRe. Because these baselines are built on SD2.1-base, we apply DTI using the same pre-trained backbone to ensure fairness. The results demonstrate that DTI attains higher text fidelity while maintaining subject similarity.

Figure 10: **Qualitative results with TI/DTI with LoRA on SDXL.** We have performed qualitative comparison of applying TI and DTI on model fine-tuning methods using LoRA (rank 32). DTI consistently improves the text prompt fidelity compared to TI.

Figure 11: **Stylization.** Qualitative comparison of personalization with diverse style inputs.

Figure 12: **My subject in my style.** Qualitative comparison of subject-style mixing within the same prompt.

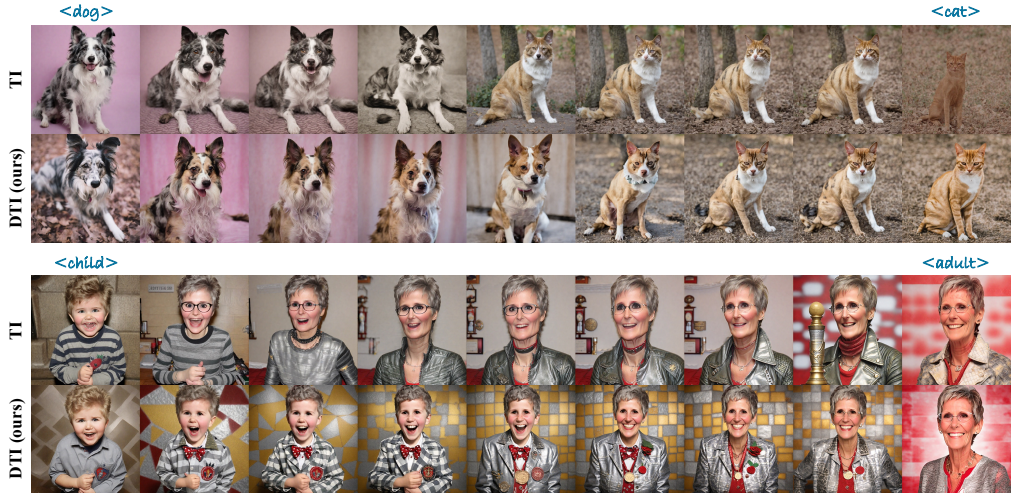


Figure 13: **Interpolation.** We compare images generated by a TI and our DTI. Two personalized subjects are interpolated, including interpolation between inanimate and animate subjects, live subjects, and human faces. Images are generated with interpolation ratio $[0.0, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 1.0]$ for better visualization. Our DTI offers smooth interpolation between concepts, expanding the personalization in more creative axis.

specific and detailed images can raise concerns related to copyright infringement, as personalized generative models may inadvertently or intentionally reproduce objects protected by intellectual property laws. Therefore, we note that it is important for users and distributors of the model to develop comprehensive awareness and implement guidelines addressing copyright boundaries, fair use, and ethical content generation. Moreover, we note that, since our method does not modify the underlying parameters of the generative model but solely adjusts the token embeddings that capture personalized concepts, the quality of generated images inherently depends on the capabilities of the underlying text-to-image model.

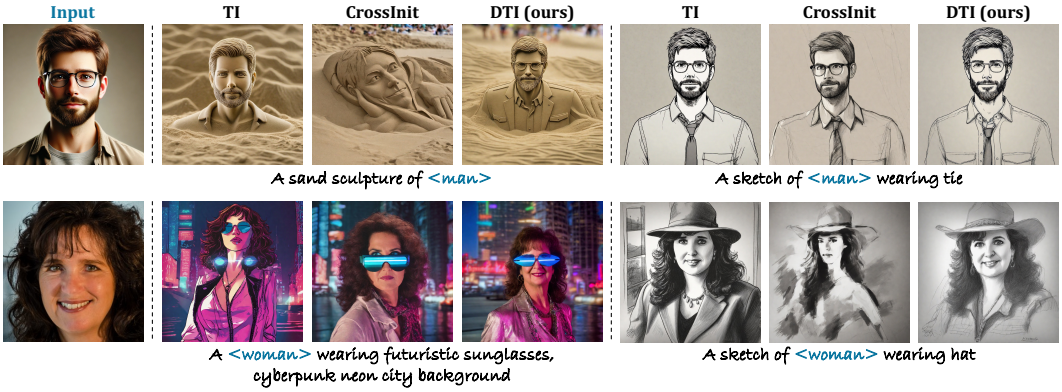


Figure 14: **Comparison of face personalization methods.** We compare our method and Textual Inversion (TI) against CrossInit, which specifically targets face personalization. To prevent bias from celebrity faces, we evaluate personalization using two alternative sources: images generated by DALL·E (Ramesh et al., 2021) (top row) and randomly selected images from the FFHQ (Karras et al., 2019) (bottom row).

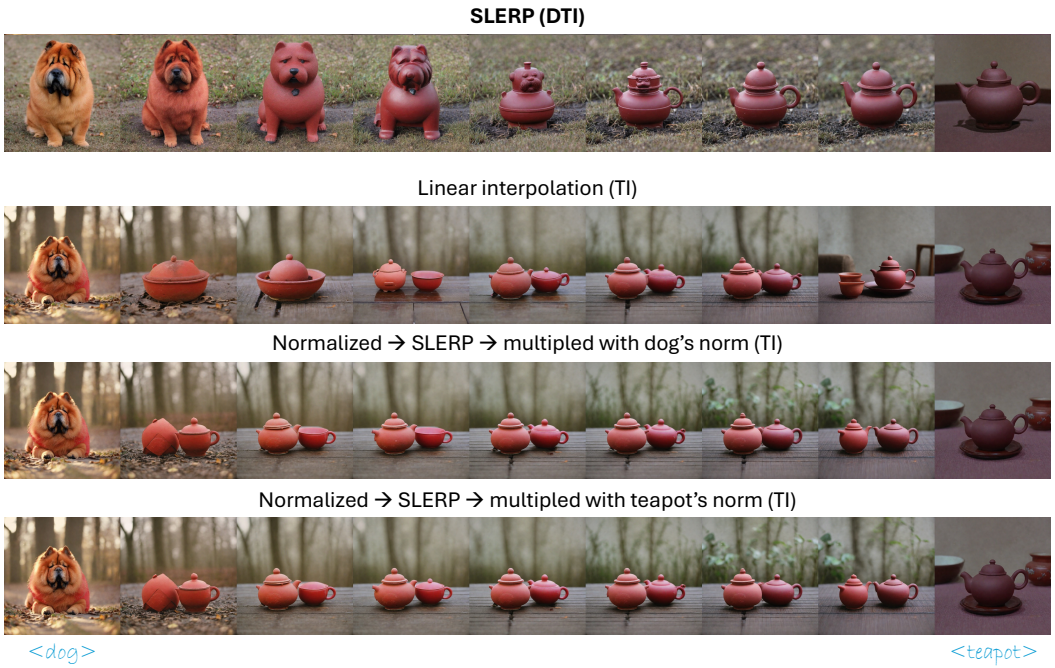


Figure 15: **Interpolation options for TI.** We compare several interpolation options for TI, including linear interpolation, SLERP with normalization and adjusted norms. While these approaches exhibit minor differences in behavior, none produce smooth transitions between concepts. In contrast, our DTI with SLERP achieves noticeably smoother and more consistent interpolations.

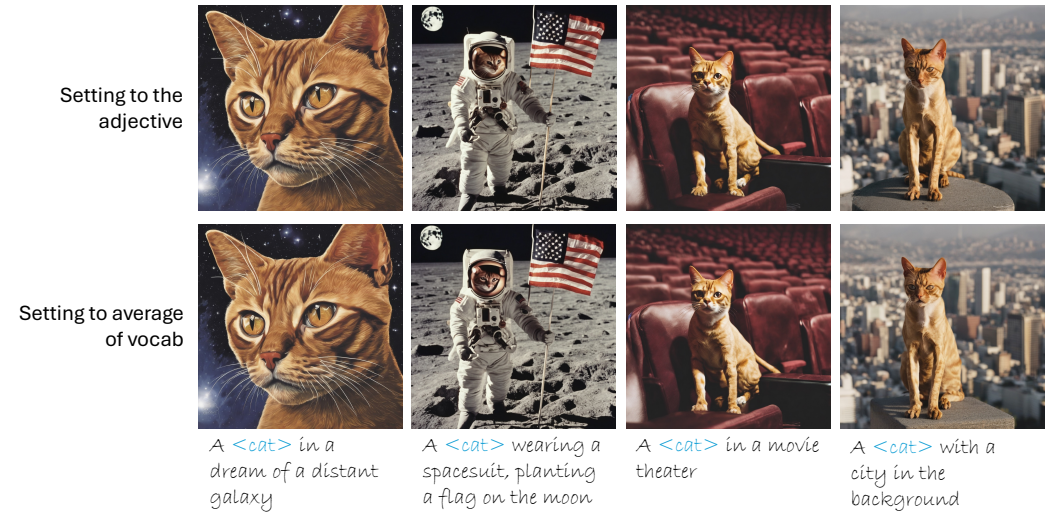


Figure 16: **Ablation on magnitude settings.** For consistency and ease of use, all magnitudes in this paper are set to the average value computed over the model's vocabulary (see ablation in Table 3). To evaluate the effect of using concept-specific magnitudes (e.g., the magnitude of 'cat' for the concept ¡cat¿), we provide ablation results under different magnitude settings. The results show that small deviations from the default magnitude do not lead to noticeable differences in output quality.

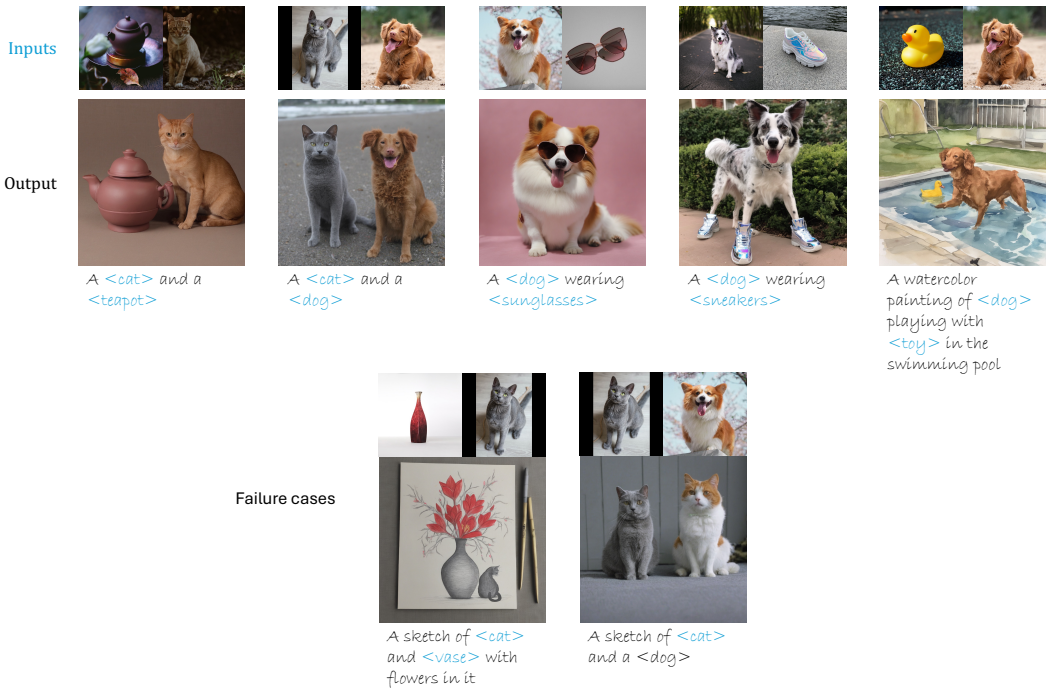


Figure 17: **Multi-concept experiments.** We further evaluate DTI by combining multiple learned concepts within a single prompt. The results demonstrate that DTI can successfully integrate multiple concepts, while the second column shows failure cases exhibiting attribute binding issues.

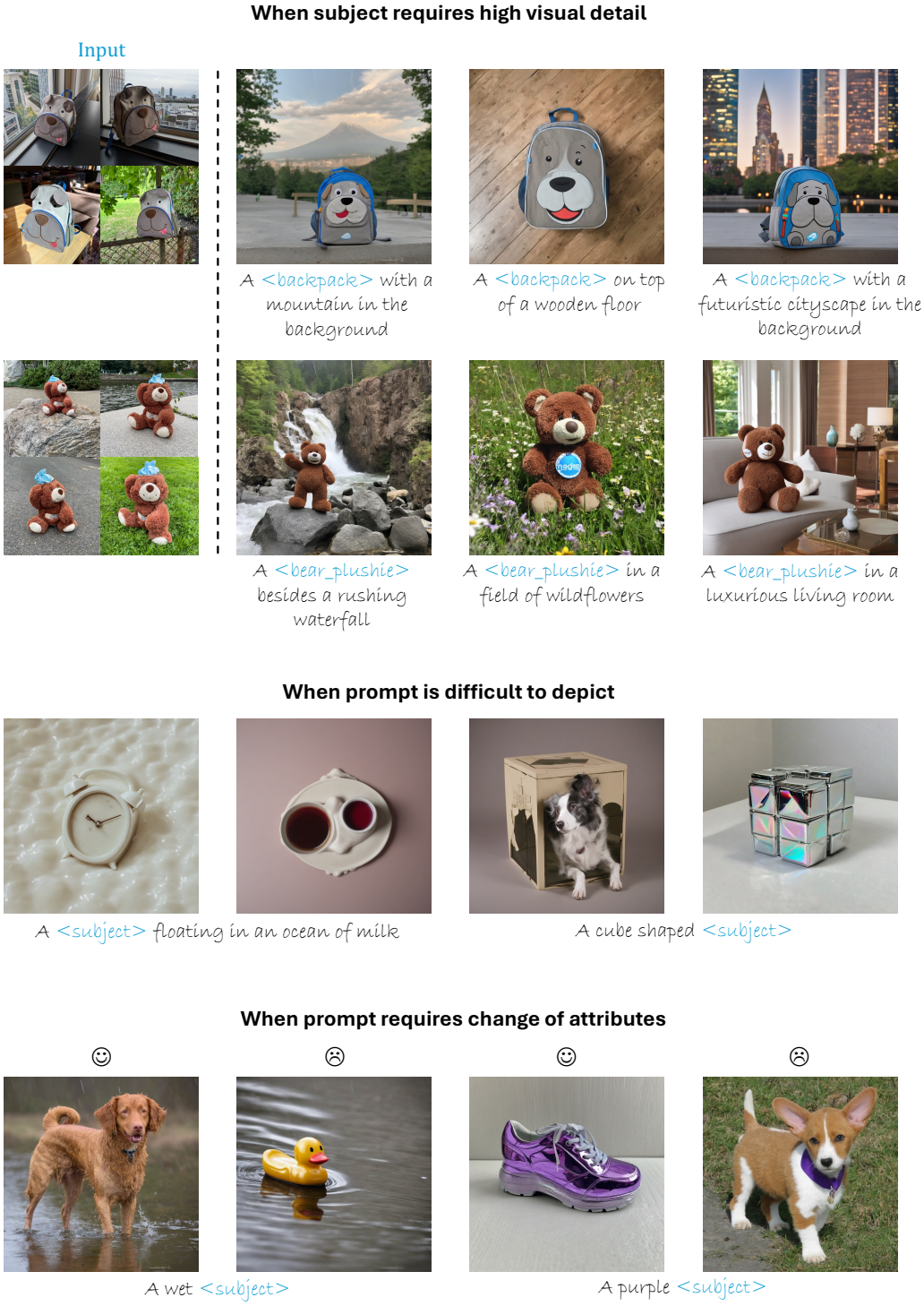


Figure 18: **Failure cases.** We present examples of three representative failure modes: (1) subjects that require high visual detail, (2) prompts that are vague or difficult to faithfully depict, and (3) prompts that involve attribute modification (e.g., color changes).