TeleStyle V2: Beyond Content-Preserving Style Transfer with Self-Distillation and Distribution-Matching-Distillation

Shiwen Zhang, **Yifan Xu**, **Haibin Huang**, **Chi Zhang**, **Xuelong Li**

TeleAI

Given a content reference and a style reference, content-preserving style transfer requires the model to generate stylized outputs with content and style consistency. We introduced TeleStyle V1 to tackle this problem. However, TeleStyle V1 is trained with photorealistic content reference and artistic style reference, which makes it incapable to cope with artistic content reference and realistic style reference in most cases. In this paper, we designed a Self-Distillation data synthesis strategy to construct such triplets from TeleStyle V1. Trained with such self-distilled triplets, our TeleStyle V2 supports Content-Style references in the forms of Realistic-and-Realistic (RnR), Realistic-and-Stylized (RnS), Stylized-and-Realistic (SnR), Stylized-and-Stylized (SnS). In addition, we found Distribution Matching Distillation could preserve the general text-guided image editing capability of the foundation model and fix the content consistency degradation caused by SFT process. Through quantitative evaluations, our TeleStyleV2-QIE-2509-DMD performs at least on par with Qwen-Image-Edit-2509-DMD, demonstrating strong general image editing skills beyond content-preserving style transfer. We observed the content/style reference order confusion problem in TeleStyle V1 and further introduced prompt enhancer to solve it. TeleStyle V2 uses Qwen-Image-Edit's VLM encoder, Qwen2.5-VL-7B, to generate content prompt and style prompt for free. TeleStyle V2 could achieve comparable style transfer performance with state-of-the-art commercial model, gemini-3-pro-image-preview.

Project: https://witcherofresearch.github.io/TeleStyleV2
Code: https://github.com/Tele-AI/TeleStyleV2

TeleAI

1 Introduction

In this technical report, we present TeleStyle V2 beyond our previous version, aiming to tackle several issues we observed in TeleStyle V1 (Zhang et al., 2026d) and further push the frontier of open-source style transfer models. We observed three major problems in the TeleStyle V1:

- *The model fails on stylized content references since TeleStyle v1 was trained with photo-realistic content references.*
- *The model fails on realistic style references since TeleStyle v1 was trained with artistic style references in most cases.*
- *The model confuses the order of content reference and style reference sometimes, returning an image almost identical with style reference.*

We introduce a self-distillation mechanism to construct Content-Style-Target triplets. The Content and Style References could be classified into 4 categories, Realistic-and-Realistic (RnR), Realistic-and-Stylized (RnS), Stylized-and-Realistic (SnR), Stylized-and-Stylized (SnS). Since such triplets are built with TeleStyle V1, this data synthesis strategy could be considered a self-distillation process.

In TeleStyle v1 and v2, we found Distribution Matching Distillation (DMD) (Yin et al., 2024; Fan et al., 2026) critical for preserving the content characteristics and helpful to avoid catastrophic forgetting. With one stone two birds, DMD prevents characteristics shifting and decreases inference steps. DMD enables TeleStyle V1 and V2 to perform as well as (if not better than) vanilla Qwen-Image-Edit series (2509,2511)

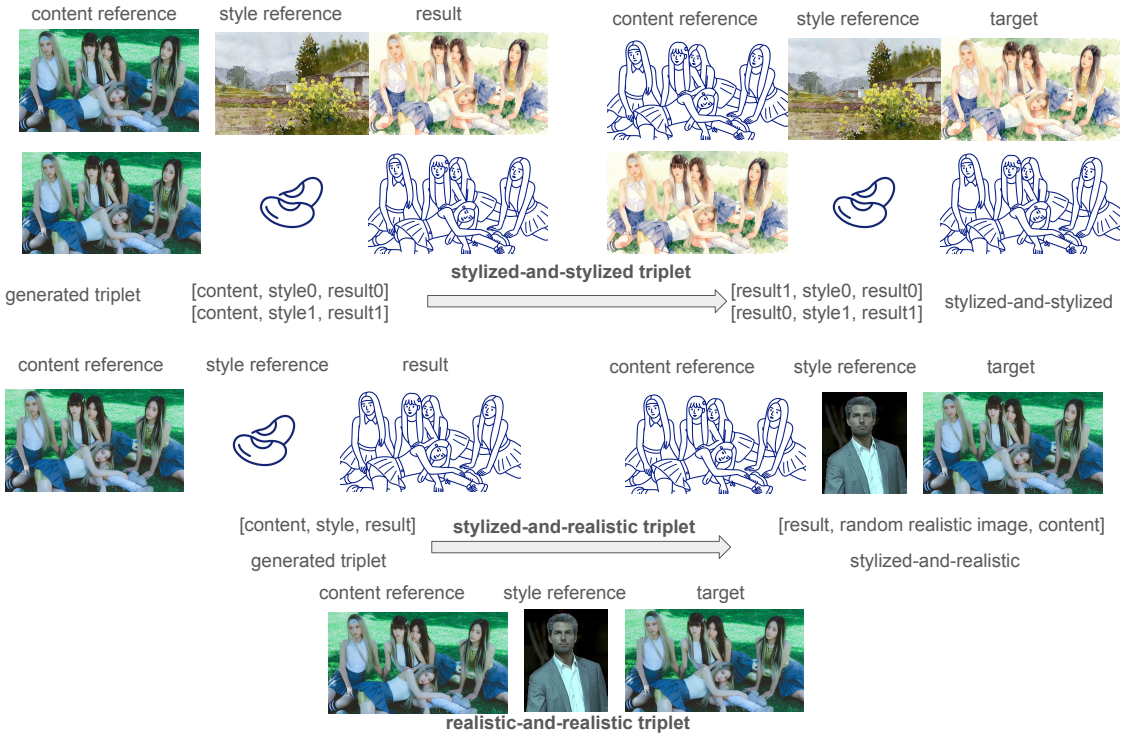

Figure 1 We construct self-distillation triplets with TeleStyle v1, to enable TeleStyle v2 coping with stylized content reference and realistic style reference, beyond realistic content reference and stylized style reference setting in TeleStyle v1. Please note that due to privacy concerns, we use idols and actors' pictures in this figure for demonstration. They **don't** exist in our training set. Our training set is built with amateurs photos, which are not shown due to privacy reasons.

(Wu et al., 2025) on GEdit-EN and GEdit-CN (Liu et al., 2025). Such a generalization capability indicates our TeleStyle could be used as a general text-guided image editing model (Kawar et al., 2023; Zhang et al., 2023; Zhang, 2024; Wu et al., 2025; Hurst et al., 2024; Team, 2025), beyond style transfer task.

In addition, we observed a phenomenon we termed *Reference Order Confusion*, where for some cases, TeleStyle generates images almost identical with style reference, showing no relation with the content reference, despite MSRoPE (Su et al., 2024; Wu et al., 2025) applied in Qwen-Image-Edit and image indices used in input template. We tried Direct Preference Optimization (DPO) (Rafailov et al., 2023) to tackle this problem, which, surprisingly, did not work. We solved this problem via prompt enhancing using VLM (Bai et al., 2025).

Our main contributions are:

- We introduced a Self-Distillation mechanism to construct Content-and-Style References of Realistic-and-Realistic (RnR), Realistic-and-Stylized (RnS), Stylized-and-Realistic (SnR), Stylized-and-Stylized (SnS). Trained with such triplets, TeleStyle V2 is strengthened in its generalization capability.
- We found utilizing Distribution-Matching-Distillation (DMD) mitigates the content consistency degradation problem and further makes TeleStyle a general text-guided image editor on par with Qwen-Image-Editing series.
- We observed the Reference Order Confusion problem and tackle it with prompt enhancement by adding a brief content description to the prompt.

For the content-preserving style transfer task, TeleStyle V2 achieves style similarity and content consistency on par with the state-of-the-art commercial model gemini-3-pro-image-preview, i.e. nano banana pro

(Team, 2025).

2 Methods

2.1 Connecting Different Style Domains via Self-Distillation

2.1.1 Brief Recap of Training Triplets in TeleStyle V1

In TeleStyle V1, we introduced how we constructed the triplet dataset [content ref, style ref, target]. The dataset is mixed with collected triplets $D_{collect}$ (Hurst et al., 2024; Song et al., 2025) and synthetic triplets $D_{synthetic}$(Li et al., 2024; Zhang et al., 2025). We proposed Style-CCL (Zhang et al., 2026b,a), a curriculum continual learning framework (Bengio et al., 2009; Zhang et al., 2020b; Zhang, 2022; Zhang et al., 2020a) to enable Qwen-Image-Edit-2509 (Wu et al., 2025) performing content-preserving style transfer task. However, in the training triplets, the content references are all realistic photographs, most of the style references are artistic images, i.e. the content reference and style reference are **R**ealistic-and-**S**tylized (RnS). Thus TeleStyle v1 is incapable of changing the style of a stylized content reference in most cases.

2.1.2 Direct Connection for Realism and Stylization

We categorize an image into binary classes, either realistic or stylized. Thus we could classify the content and style references into 4 categories: Realistic-and-Realistic (RnR), Realistic-and-Stylized (RnS), Stylized-and-Realistic (SnR), Stylized-and-Stylized (SnS). In TeleStyle v1, we only considered the RnS setting during dataset construction. In Figure 1, we show that it is quite easy to construct RnR triplets by using the same realistic image as content reference and target, then assigning a random realistic image as style reference. However, for SnS and SnR triplets, it is not very straight-forward. Since TeleStyle v1 is already a state-of-the-art content-preserving style transfer model, we introduce a self-distillation mechanism to construct SnS and SnR triplets.

For SnS, demonstrated in Figure 1, given the same content reference, we use two random style references to generate two targets, with the order of *[content ref, style ref, target]*:

$$[content, style_0, result_0] \longrightarrow [result_1, style_0, result_0], \tag{1}$$

$$[content, style_1, result_1] \longrightarrow [result_0, style_1, result_1]. \tag{2}$$

Thus, we obtain two triplets, where the content reference, style references, and the targets, are all stylized images.

For SnR, in Figure 1, we have

$$[content, style, result] \longrightarrow [result, a\ random\ realistic\ image, content]. \tag{3}$$

The construction of training triplets utilized the existing capability itself, thus this is a self-distillation approach. Through further experiments, we found the potential of further data mining, which we leave for future exploration. Via self-distillation, we bridge and connect the disjoint realistic and stylized modalities. We merge these new triplets into our training set and train TeleStyle v2 with Style-CCL. However, we found that the SnR task is very challenging for reference-guided generation and the performance is very instable. Thus we recommend directly using the prompt "convert Figure 1 to a photorealistic photograph" on the content reference instead of using the content-preserving style transfer mode.

2.2 Distribution Matching Distillation is the Savior

The training of TeleStyle leads to certain content consistency degradation and image distortion to the Qwen-Image-Edit foundation models, shown in Figure 2. Though this is far from "catastrophic" forgetting, we could still fix such degradation via minimizing the reverse Kullback-Leibler (KL) Divergence $D_{KL}(p_{fake} \| p_{real})$

(Kullback and Leibler, 1951). With the real data distribution $p_{real}(x_0)$ and the generated data distribution $p_{fake}(x_0)$ produced by $\boldsymbol{G}_{\boldsymbol{\phi}}$, the gradient of reverse KL Divergence is:

$$\nabla_{\boldsymbol{\phi}} D_{KL}(p_{fake} \| p_{real}) = E_{z, x_0 = \boldsymbol{G}_{\boldsymbol{\phi}}(z)}[(\nabla_{x_0} \log p_{fake}(x_0) - \nabla_{x_0} \log p_{real}(x_0)) \frac{d\boldsymbol{G}}{d\boldsymbol{\phi}}] \tag{4}$$

where $z \sim \mathcal{N}(0, \boldsymbol{I})$ is a random Gaussian noise input.

With Distribution Matching Distillation (DMD) (Yin et al., 2024; Fan et al., 2026; Contributors, 2025) models on Qwen-Image-Series applied to TeleStyle, we could hit two birds with one stone. The inference cost is reduced 10 times, and the content consistency and vanilla image editing capability of Qwen-Image-Editing series are preserved, even strengthened. TeleStyle could conduct content-preserving style transfer and text-guided image editing in one run (the first example in Figure 2). In addition, TeleStylev2 accepts optional number of reference images. For example, it works on style reference + prompt scenario by providing only one style reference. It also works on traditional text-guided image editing task by providing one content reference. We also compare TeleStyle v2 with and without DMD in Figure 2. TeleStyle v2 works without DMD, though the aesthetics could drop and the image could be distorted sometimes. It is worth noting that TeleStyle v2 boosts Qwen-Image-Edit series on style-related tasks for text-guided image editing, even without being trained on this text-guided task.

2.3 Tackling Reference Order Confusion

We observed TeleStyle V1 sometimes confuses the order of content and style reference in a certain way so that the generated result is very similar to the style reference and has no relation with content reference. We term such a phenomenon *Reference Order Confusion*. Though Qwen-Image-Edit series have already applied MS-RoPE to distinguish the references and further encoded image indices in the VLM input template, the problem persists for some cases. To tackle the Reference Order Confusion, We tried Direct Preference Optimization (DPO) (Rafailov et al., 2023), and surprisingly found it not working. We finally solve this problem by adding auto-generated content description to the prompt. Such a simple hint inspires the model to understand the actual target.

2.3.1 Direct Preference Optimization Does Not Work

We first tried Direct Preference Optimization (DPO) to mitigate the reference order confusion. We tried two data scales of win-lose training pairs $[x_0^{win}, x_0^{lose}]$: 2k and 400k. The 2k data is obtained by manually filtering win-lose pairs from a 100k fraction of 10 million generated results of TeleStyle V1. The 400k data is constructed from triplet data of section 2.1, where we use target image as the winning image and style reference as the lost image. The loss of DPO is:

$$\begin{cases} \text{Diff}_{\text{policy}} = \left(\left\| v_\theta(x_t^{win}, h, t) - v_t^{win} \right\|_2^2 - \left\| v_\theta(x_t^{lose}, h, t) - v_t^{lose} \right\|_2^2 \right) \\ \text{Diff}_{\text{ref}} = \left(\left\| v_{\text{ref}}(x_t^{win}, h, t) - v_t^{win} \right\|_2^2 - \left\| v_{\text{ref}}(x_t^{lose}, h, t) - v_t^{lose} \right\|_2^2 \right) \\ \mathcal{L}_{DPO} = -\mathbb{E}_{h, (x_0^{win}, x_0^{lose}) \sim \mathcal{D},\, t \sim \mathcal{U}(0,1)} \left[\log \sigma \Big(-\beta \left(\text{Diff}_{\text{policy}} - \text{Diff}_{\text{ref}} \right) \Big) \right], \end{cases} \tag{5}$$

where $\text{Diff}_{\text{policy}}$ and Diff_{ref} denote the preference differences computed by the policy model v_θ and the reference model v_{ref}, respectively, β is a scaling parameter, and $\sigma(\cdot)$ denotes the sigmoid function. h denotes the vision-text embedding from VLM, x_0^{win} denotes the VAE (Kingma and Welling, 2014) latent of winning image, and x_0^{lose} for the losing image. Timestep is sampled from $t \sim (0, 1)$ to construct the input latent variable x_t^{win} and x_t^{lose} as well as their corresponding velocity v_t^{win} and v_t^{lose}.

We trained a DPO Lora on Qwen-Image-Edit-2509. Unfortunately and quite surprisingly, DPO could not solve this reference order confusion problem, not even with 400K data.

2.3.2 Prompt Enhancement: The Solution

With further investigating, we found a very simple solution: with a very brief description of the objects in the content reference with a few separated words, the model immediately understands the editing purpose.

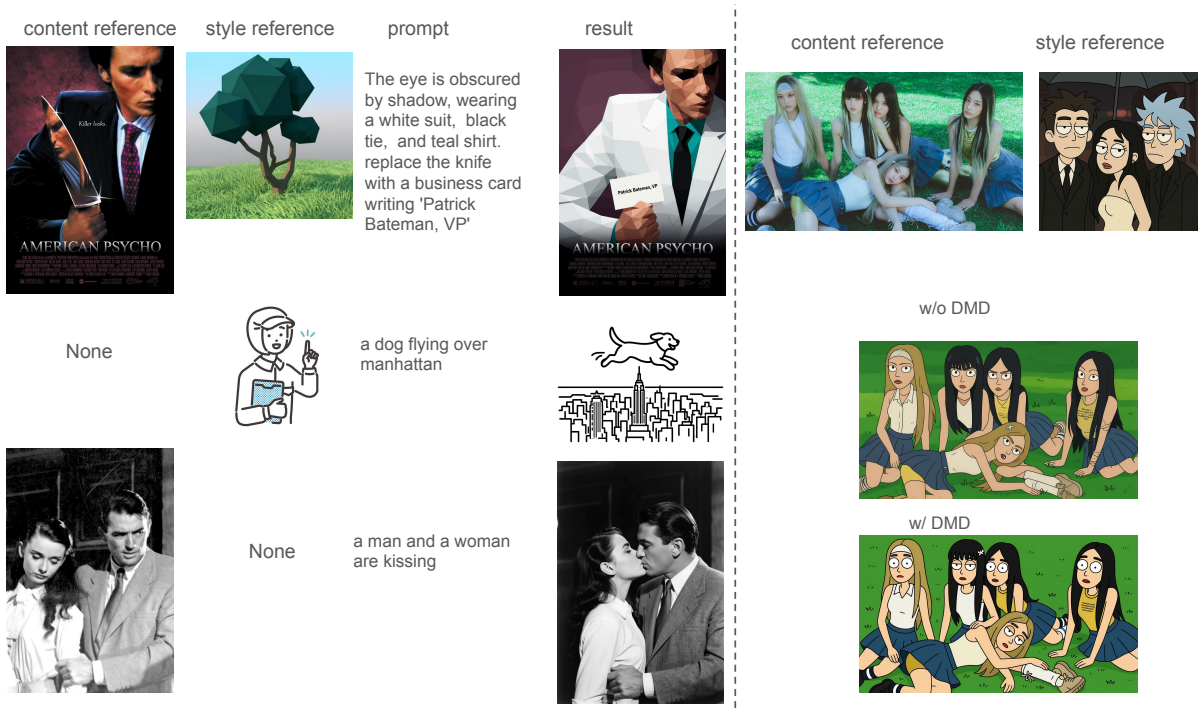

Figure 2 DMD helps TeleStyle to preserve the general image editing capability of Qwen-Image-Edit series and mitigate the degradation caused by style transfer SFT.

Thus we use the Qwen2.5-VL-7B (Bai et al., 2025) to complete this task, which has been already loaded into the memory as the visual-language encoder of Qwen-Image-Edit model. The generated words are called content prompt. We further use Qwen2.5-VL-7B to describe the style reference, which is called style prompt. The content and style prompts are optional during inference, with the content prompt solving the Reference Order Confusion problem and style prompt might solving some corner cases to enhance style similarity.

3 Experiments and Evaluation

3.1 Implementation

We trained TeleStyle v2 on Qwen-Image-Edit-2509, Qwen-Image-Edit-2511 (Wu et al., 2025) and FLUX 2 Klein (Labs, 2025), which are foundation models based on mmdit structure (Peebles and Xie, 2023; Esser et al., 2024; Labs, 2024, 2025) instead of UNet (Ho et al., 2020; Podell et al., 2023). We trained both Lora (Hu et al., 2021) and full-parameter models (Wu et al., 2025) on these foundation models. However, all the qualitative and quantitative results shown in this paper are produced by TeleStyleV2-QIE-2509 Lora. The optimizer is AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2017), and the learning rate is 1e-4. The Loras are trained on 4 H100 with gradient checkpointing (Griewank and Walther, 2000) and full-parameter models on 8 H100 with DeepSpeed (Rasley et al., 2020).

3.2 Evaluation on General Image Editing Capability

We test the general image editing capability of TeleStyle v2 on GEdit-Bench (Liu et al., 2025) evaluated by Qwen2.5-VL-72B (Bai et al., 2025). Shown in Table 1, with Qwen-Image-Edit-2509 (QIE-2509) as the base model, TeleStylev2-DMD demonstrates comparable text-guided image editing performance with QIE-2509-DMD on 11 different sub-tasks. Such metrics prove that beyond strong content-preserving style transfer capability, TeleStylev2-DMD also demonstrates competitive text-guided image editing performance on par with vanilla Qwen-Image-Edit. Though our training triplets are all English, TeleStylev2 could even out-

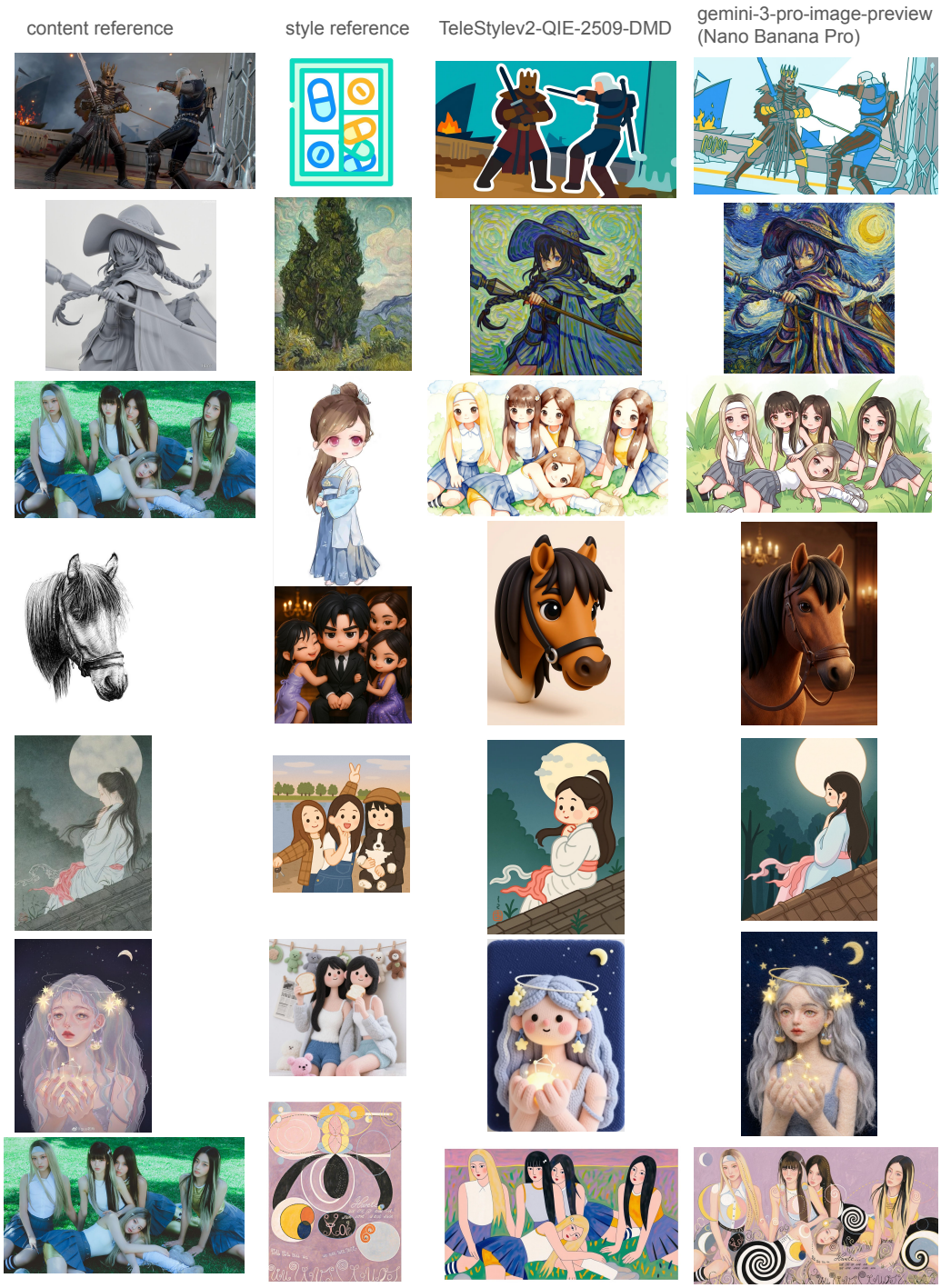

Figure 3 Qualitative Comparison of TeleStylev2-QIE-2509-DMD with the state-of-the-art image generation model, gemini-3-pro-image-preview (nano-banana-pro).

perform vanilla QIE-2509-DMD on GEdit-Chinese. Specifically, there is a noticeable improvement on the style-changing sub-task, though we did not train TeleStylev2 on any text-guided style transfer data.

Model	GEdit-EN (Full set) ↑						GEdit-CN (Full set) ↑					
	QIE-2509-DMD			TeleStylev2-QIE-2509-DMD			QIE-2509-DMD			TeleStylev2-QIE-2509-DMD		
	Q_SC	**Q_PQ**	**Q_O**	**Q_SC**	**Q_PQ**	**Q_O**	**Q_SC**	**Q_PQ**	**Q_O**	**Q_SC**	**Q_PQ**	**Q_O**
background change	8.050	7.700	7.866	8.150	7.650	7.836	8.000	7.700	7.838	8.050	7.675	7.800
color alter	8.500	7.300	7.843	8.225	7.250	7.570	8.275	7.225	7.683	8.275	7.275	7.709
material alter	6.850	7.025	6.386	7.075	6.975	6.538	6.800	6.725	6.259	6.825	7.025	6.347
motion change	7.425	7.750	7.467	7.550	7.800	7.545	7.475	7.650	7.420	7.325	7.775	7.326
ps human	6.714	7.529	6.711	6.571	7.543	6.539	6.500	7.543	6.555	6.543	7.600	6.618
style change	6.583	7.183	6.479	7.133	7.083	**6.943**	6.867	7.167	6.606	7.267	7.083	**7.053**
subject add	7.950	7.667	7.720	7.967	7.617	7.697	8.050	7.583	7.761	7.983	7.717	7.832
subject remove	8.263	7.667	7.803	8.123	7.719	7.752	7.561	7.667	7.185	8.158	7.754	7.794
subject replace	8.317	7.483	7.865	8.217	7.567	7.803	8.083	7.483	7.691	8.150	7.533	7.816
text change	9.051	7.525	8.225	9.020	7.505	8.157	8.747	7.485	8.013	8.859	7.505	8.121
tone transfer	7.150	7.400	7.136	7.075	7.500	6.978	7.075	7.325	6.886	7.400	7.350	7.284
avg	7.714	7.475	7.409	7.737	7.474	7.396	7.585	7.414	7.263	7.712	7.481	7.427

Table 1 Quantitative evaluation on GEdit-EN and GEdit-CN. We evaluate Qwen-Image-Edit-2509-DMD (QIE-2509-DMD) and TeleStylev2-QIE-2509-DMD from all aspects of GEdit-English and GEdit-Chinese instructions. TeleStylev2 performs on par with, or even better than QIE-2509-DMD on such general text-guided image editing tasks, with specific improvement on the style-related task. Semantic Consistency (Q_SC), Perceptual Quality (Q_PQ), and Overall Score (Q_O) refer to the metrics evaluated by Qwen2.5-VL-72B.

3.3 Comparison with gemini-3-pro-image-preview

We compare TeleStylev2-QIE-2509-DMD with the state-of-the-art image generation model gemini-3-pro-image-preview (nano banana pro) (Team, 2025) in Figure 3. Our model outperforms gemini-3-pro-image-preview in terms of style similarity in most cases and on par with it in terms of content consistency. We choose not to release the quantitative comparison and user study because we found TeleStyle V2's win rate against nano-banana-pro is very high, which is quite unexpected and we suspect that our testing set might be biased. To make a safe statement, TeleStyleV2-QIE2509-DMD almost performs on par with gemini-3-pro-image-preview on content-preserving style transfer task.

4 Conclusion

In this work, we introduce TeleStyle V2, improving the content-preserving style transfer performance beyond TeleStyle V1 (QwenStyle). TeleStyle V2 introduces self-distillation to improve generalization and employs distribution-matching-distillation to preserve the general image editing capability of foundation models and to repair the content consistency degradation caused by SFT. Prompt Enhancer instead of DPO solves the reference order confusion problem and further solves some corner cases.

The training of TeleStyle V2 was finished in March, 2026. In the near future, we will release a multi-reference subject-style driven Omnimodel, based on self-distillation of TeleStyle and TeleComposer (Zhang et al., 2026c).

References

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.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

LightX2V Contributors. Lightx2v: Light video generation inference framework. https://github.com/ModelTC/lightx2v, 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 *Forty-first international conference on machine learning*, 2024.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41667–41676, 2026.

Andreas Griewank and Andrea Walther. Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Transactions on Mathematical Software (TOMS)*, 26(1):19–45, 2000.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017. IEEE, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Gemini Team. Gemini-3-pro-image-preview. 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Shiwen Zhang. Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning. *arXiv preprint arXiv:2203.05928*, 2022.

Shiwen Zhang. Fast Imagic: Solving Overfitting in Text-guided Image Editing via Disentangled UNet with Forgetting Mechanism and Unified Vision-Language Optimization. In *PMLR*, 2024.

Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R Scott, and Limin Wang. V4d: 4d convolutional neural networks for video-level representation learning. In *International Conference on Learning Representations*, 2020a.

Shiwen Zhang, Sheng Guo, Limin Wang, Weilin Huang, and Matthew Scott. Knowledge integration networks for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020b.

Shiwen Zhang, Shuai Xiao, and Weilin Huang. Forgedit: Text guided image editing via learning and forgetting. *arXiv preprint arXiv:2309.10556*, 2023.

Shiwen Zhang, Zhuowei Chen, Lang Chen, and Yanze Wu. Cdst: Color disentangled style transfer for universal style reference customization. *arXiv preprint arXiv:2506.13770*, 2025.

Shiwen Zhang, Haibin Huang, Chi Zhang, and Xuelong Li. Qwenstyle: Content-preserving style transfer with qwen-image-edit. *arXiv preprint arXiv:2601.06202*, 2026a.

Shiwen Zhang, Haoyuan Wang, Xianghao Zang, Haibin Huang, Chi Zhang, and Xuelong Li. Style-ccl: Content-preserving style transfer via curriculum continual learning. *arXiv preprint arXiv:2606*, 2026b.

Shiwen Zhang, Yifan Xu, Haibin Huang, Chi Zhang, and Xuelong Li. Telecomposer: Multi-reference-driven image and video customization. *arXiv preprint arXiv:2606*, 2026c.

Shiwen Zhang, Xiaoyan Yang, Bojia Zi, Haibin Huang, Chi Zhang, and Xuelong Li. Telestyle: Content-preserving style transfer in images and videos. *arXiv preprint arXiv:2601.20175*, 2026d.