Unified Customized Generation by Disentangled Reward Modeling

Shaojin Wu[1] Mengqi Huang[1,2*] Yufeng Cheng[1] Wenxu Wu[1]
Jiahe Tian[1] Yiming Luo[1] Fei Ding[1†] Qian He[1]
[1]Intelligent Creation Team, ByteDance, [2]University of Science and Technology of China
{wushaojin, chengyufeng.cb1, tianjiahe.00, luoyiming.lym, dingfei.212, heqian}@bytedance.com

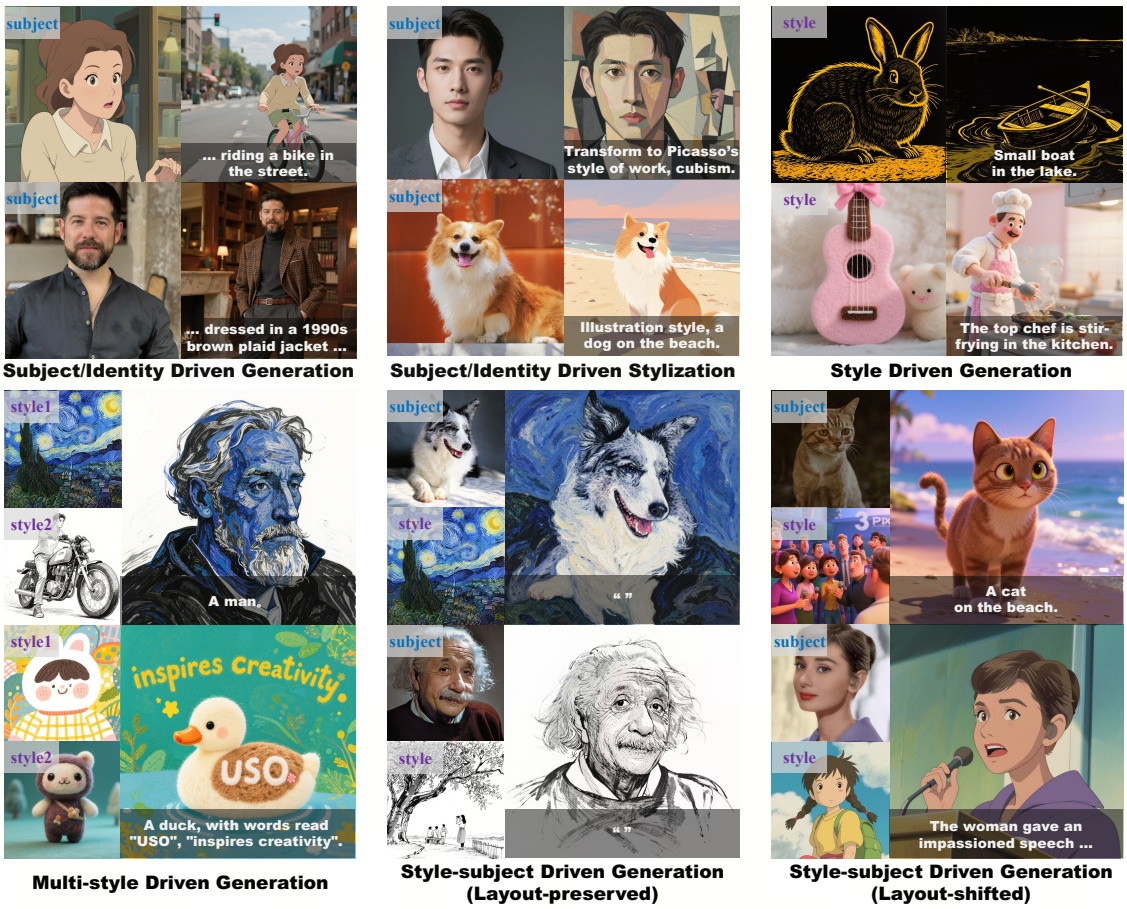


Figure 1. We propose **USO**, a unified model that jointly optimizes for subject and style, enabling customizable generation with high subject consistency and style fidelity.

Abstract

*Existing literature typically treats various customized generation tasks (e.g., subject-customized generation, style-customized generation) as distinct and disjoint problems, with each task focusing solely on customizing a specific aspect of the reference image. However, we argue that the objectives of these different customization tasks are inherently complementary and can be mutually enhanced within a unified framework, as they fundamentally involve the disentanglement of multiple feature aspects from the reference image. To this end, we introduce **USO**, a **U**nified **S**imultaneous **O**ptimization framework to simultaneously unify different customized tasks (i.e., subject and style). Specifically, USO introduces a cyclical data-model framework that connects these two tasks by a subject-for-style*

∗ Corresponding author. Email: huangmq@ustc.edu.cn.
† Project lead

data curation pipeline and a style-for-subject model training pipeline. The subject-for-style data curation pipeline leverages a state-of-the-art subject-customized model to generate high-quality triplet data comprising content images, style images, and their corresponding stylized content images. Building on this foundation, the style-for-subject model training pipeline introduces an auxiliary style reward to simultaneously align style and content features, thereby reinforcing the model's ability to extract the desired style or content features from the reference image. Extensive experiments demonstrate that USO achieves state-of-the-art performance among open-source models, excelling in both subject consistency and style similarity. Code and model: https://github.com/bytedance/USO.

1. Introduction

The significant advancements in image generation [15, 22, 24, 27, 29] over the past years have greatly improved generative controllability, fundamentally changing how humans create images, *i.e.*, whether through abstract textual descriptions, specific visual reference images, or both. Research on leveraging both textual and visual conditions has attracted increasing interest, giving rise to numerous real-world tasks such as style-driven generation and subject-driven generation. While textual conditions are typically explicit, ***visual conditions are inherently noisy***, as images intrinsically embody a rich spectrum of features (*e.g.*, style, appearance), of which only a specific one is relevant to a specific task. For instance, style-driven generation requires only the style feature from the reference images, whereas other features constitute noise. Therefore, a fundamental and long-standing challenge in these tasks is to accurately ***include all required features from the reference image while simultaneously excluding other noisy ones***, *e.g.*, including only the style in style-driven generation or only the subject's appearance in subject-driven generation.

Extensive efforts in the literature have been dedicated to disentangling different features in visual conditions (*i.e.*, reference images). On the one hand, in the realm of style-driven generation, DEADiff [26] employs QFormer to selectively query only the style features from reference images. CSGO [40] constructs content-style-stylized triplets to facilitate style-content decoupling during training. StyleStudio [17] introduces style-based classifier-free guidance (SCFG) to enable selective control over stylistic elements and to mitigate the influence of irrelevant features. On the other hand, subject-driven generation methods primarily focus on disentangling subject appearance features or constructing more effectively disentangled paired data. For example, RealCustom [12, 20, 21] proposes a dual-inference framework that selectively incorporates subject features into subject-specific regions. UNO [37] leverages the in-context capabilities of DiT to progressively improve both the quality of paired data and the model itself. To conclude, existing methods primarily focus on ***task-specific disentanglement*** by designing tailored datasets or model architectures for each individual task, thereby performing ***disentanglement in an isolated, single-task context.***

We argue that a more comprehensive and precise disentanglement approach should fully account for the coupling and complementarity between different generation tasks. Each task should not only learn which features to include, but, more importantly, also learn which features to exclude, *i.e.*, features that are often required by other tasks. ***Therefore, learning to include certain features in one task inherently informs and enhances the process of learning to exclude those same features in a complementary task, and vice versa.*** For example, style-driven generation aims to incorporate stylistic features while excluding subject appearance features, whereas subject-driven generation does the exact opposite. The ability to learn and include subject appearance features in subject-driven generation can, in turn, help style-driven generation more effectively learn to exclude these features, thereby improving disentanglement for both tasks. In conclusion, we believe that jointly modeling complementary tasks enables a mutually reinforcing disentanglement process, leading to a more precise separation of relevant and irrelevant features for each task.

Based on the above motivation, we propose a novel ***cross-task co-disentanglement*** paradigm to unify subject-driven and style-driven generation, and, more importantly, to mutually enhance the performance of both tasks. Specifically, this co-disentanglement paradigm is implemented through a *subject-for-style* data curation framework and a *style-for-subject* model training framework. The *subject-for-style* framework first utilizes a state-of-the-art subject model to generate high-quality style data, while the *style-for-subject* framework subsequently trains a more effective subject model under the guidance of style rewards and disentangled training. Technically, on the one hand, for the *subject-for-style* data curation framework, we build upon a state-of-the-art subject-driven model [16, 37] and further develop both a stylization expert and a de-stylization expert to curate stylized and non-stylized images. This process ultimately constructs triplet data pairs in the form of <style reference, de-stylized subject reference, stylized subject result> for subsequent model training. On the other hand, for the *style-for-subject* model training framework, we propose a **U**nified **S**imultaneous **O**ptimization framework, which introduces task disentanglement training and auxiliary style reward.

Our contributions are summarized as follows:

Concept: We point out that existing style-driven and subject-driven methods focus solely on isolated disentanglement within each task, neglecting their potential com-

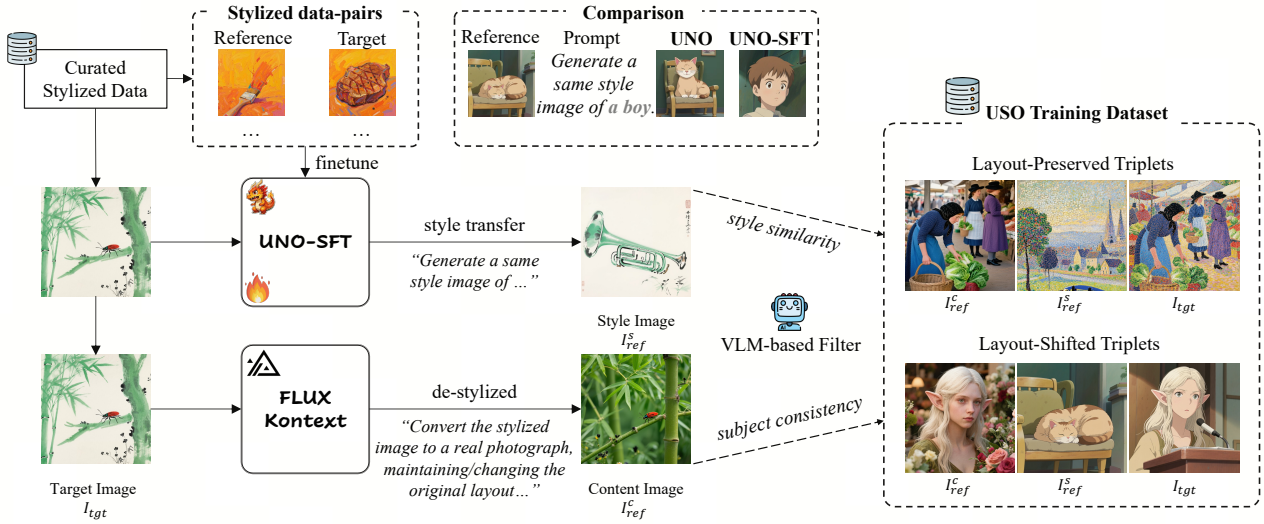


Figure 2. Illustration of our proposed cross-task triplet curation framework, which systematically generates layout-preserved and layout-shifted triplets.

plementarity and thus leading to suboptimal disentanglement. For the first time, we propose a novel cross-task co-disentanglement paradigm that unifies style-driven and subject-driven tasks, enabling mutual enhancement and achieving significant performance improvements for both.

Methodology: We present a novel cross-task triplet curation framework that bridges style-driven and subject-driven generation. Building on this, we introduce USO, a unified customization architecture that incorporates task disentanglement training and auxiliary style rewardto further promote cross-task disentanglement. We further release USO-Bench, to the best of our knowledge, the first benchmark tailored for evaluating cross-task customization.

Performance: Extensive evaluations on USO-Bench and DreamBench [28] show that USO achieves state-of-the-art results on subject-driven, style-driven, and joint style-subject-driven tasks, attaining the highest CLIP-T, DINO, and CSD scores. USO can handle individual tasks and their free-form combinations while exhibiting superior subject consistency, style fidelity, and text controllability as shown in Figure 1.

2. Related Work

2.1. Subject-Customized Generation

Subject-customized generation refers to generating images of the same subject conditioned on a text instruction and reference images of given subjects. Dreambooth [28] and IP-Adapter [42] turn a UNet-based text-to-image model into a subject-driven model by parameter-efficient tuning or a newly introduced attention plug-in. Recently, popular image-generation foundation models have shifted from UNet-based architectures to transformer-based ones. The inherent in-context learning capabilities of transformers have greatly enriched research on subject-driven generation. ICLoRA [11], OminiControl [31], UNO [37], UMO [2], and FLUX.1 Kontext [16] use shared attention between the generated image and reference image to train a text-to-image DiT into a subject-driven variant. It is worth noting that some of them have extended the reference subject to other types. OminiControl [31] supports layout control image as a reference, UNO [37] supports multiple reference images input, and DreamO [23] can work for simple style transfer. They have indicated that various types of reference-guided generation can be unified within the DiT in-context framework. This further prompts the question of whether jointly addressing different tasks in this setting could lead to mutual benefits across them.

2.2. Style-Customized Generation

Style-customized generation aims to apply the style in the reference image to the given content image or fully generated image. Early work like adaptive instance normalization [13] achieved impressive style transfer results with layout-preserved results by simply using a pre-trained network as the style encoder and well-designed injection modules. The recent powerful text-to-image generation base models, like Stable Diffusion [5, 24] and FLUX [15], along with style transfer plugins built upon them, have significantly improved the convenience and effectiveness of performing this task. Several are even training-free, like StyleAlign [38] and StylePrompt [14] which transfer the style via simple query-key swapping in the specific self-attention layers. Other training-based methods can theoretically achieve better fitting and style transfer performance, but they also raise concerns of content leakage. IP-Adapter [42] and DEADiff [26] demonstrate the style transfer ability with a new decoupled cross-attention layer

trained with coupled data, and overcome the content leakage by decreasing the injection weights in inference-time. InstanceStyle [32], StyleShot [8] and B-lora [6] provide more detailed time-aware and layer-aware injection strategies to disentangle the style and content feature injections. However, those disentanglement analyses are tied to the specific model architecture and hard to migrate.

3. Methodology

3.1. Cross-Task Triplet Curation Framework

This section details the construction of cross-task triplets for USO training. Although prior works [33, 40] have explored triplet generation, they retain the original layout, preventing any pose or spatial re-arrangement of the subject. To jointly enable subject-driven and style-driven generation beyond simple instruction-based edits, we curate a new USO dataset expressly designed for this unified objective.

Figure 2 provides an overview of USO dataset. Our co-disentanglement paradigm starts from a *subject-for-style* data curation framework. Among many possible tasks, subject-driven (i.e., UNO-1M [37]) and instruction-based editing (i.e., X2I2 [36]) datasets are comparatively easy to collect at scale, enabling targeted task-specific corpora. In particular, subject-driven data emphasizes learning from content cues while preserving subject identity and consistency; instruction-based editing bridges styles by preserving spatial layout and transferring appearance between realistic and stylized domains in both directions. These resources naturally support training domain-specialist models and, through deliberate dataset design, induce the capabilities we care about (*e.g.*, extracting task-relevant features conditioned on image type). Guided by these insights, we curate $200k$ stylized image pairs sourced from publicly licensed datasets and augmented with samples synthesized by state-of-the-art text-to-image models.

Then, we formulated two complementary experts on top of the leading customization frameworks UNO [37] and FLUX.1 Kontext dev [16]:

Stylized expert model. We fine-tune a UNO-SFT model using curated stylized data-pairs. This enables style-driven generation conditioned on a style-reference image, producing a new subject rendered in the target style (I^s_{ref} from I_{tgt}). As shown in Figure 2, after supervised finetuning, after supervised fine-tuning, UNO-SFT produces results of high style similarity without content leakage. These results are further refined using VLM.

De-stylization expert model. To implement this, we leverage the frozen FLUX.1 Kontext dev for its powerful instruction editing capabilities. This allows inversion of stylized images into photorealistic counterparts, supporting either flexible layout shifts or preservation (I^c_{ref} from I_{tgt}).

Each curated stylized image serves as the target I_{tgt}. We synthesize its style reference I^s_{ref} via the stylization expert and its content reference I^c_{ref} via the de-stylization expert. Following [37], a VLM-based filter enforces style similarity between I_{tgt} and I^s_{ref} and subject consistency between I_{tgt} and I^c_{ref}. This yields two kinds of triplets, shown in Figure 2: layout-preserved and layout-shifted. Unlike prior work [33, 40], which focuses solely on style-driven generation and confines itself to layout-preserved triplets, our cross-task triplets achieve deeper content–style disentanglement across tasks and are used to train USO.

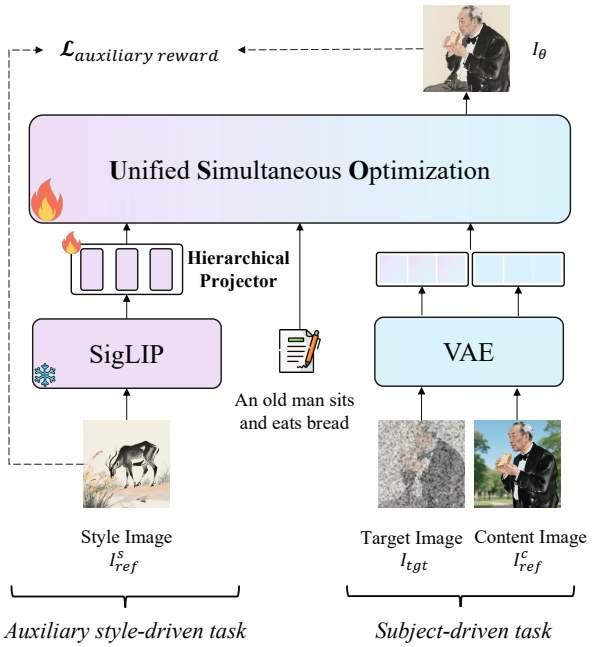


Figure 3. Illustration of the training framework of **USO**.

3.2. Unified Customization Framework (USO)

To learn how different customization tasks complement and mutually enhance each other within a unified framework, we treat the subject-driven task as the primary task and the style-driven task as auxiliary. We then unify the two tasks using a single model. We train USO on two kinds of triplets from Section 3.1. Critically, in addition to layout-preserved triplets, we introduce layout-shifted triplets, which involve changes in spatial configuration. These are essential for building a robust unified model because they force the network to inject desired stylistic features while maintaining subject consistency across diverse stylized scenarios and various text prompts.

3.2.1. Task Disentanglement Training.

Disentangled conditional encoder. As illustrated in Figure 3, We start from a pre-trained text-to-image (T2I) model and fine-tune it into a text-image-to-image (TI2I) model.

We use different encoders to process different types of conditioned images. For the style-driven task, we employ

Figure 4. Qualitative comparison with different methods on subject-driven generation.

the semantic encoder SigLIP [43] to process the reference style image I^s_{ref}. While subject-driven or identity-preserving tasks typically emphasize high-level semantics, style-driven tasks must simultaneously handle two extremes: high-level semantics to accommodate large geometric deformations (e.g., 3-D cartoon styles) and low-level details to reproduce subtle brushstrokes (e.g., pencil sketches). Following recent works like [9, 44], we introduce a lightweight Hierarchical Projector $\mathcal{M}_{\text{Proj}}(\cdot)$ to project multi-scale, fine-grained visual features z_s from the extracted SigLIP embeddings $\{c_i\}_{i=1}^N$, where N represents the layer indices of SigLIP. This process can be formulated as:

$$z_s = \text{Concatenate}(\mathcal{M}_{\text{Proj}}(\{c_i\}_{i=1}^N)), \tag{1}$$

Then we introduce subject conditioning as shown in Figure 3. Following recent paradigms [31, 37], the content image I^c_{ref} is encoded into pure conditional tokens z_c by a frozen VAE encoder $\mathcal{E}(\cdot)$. We formulate USO as a multi-image conditioned model, yet explicitly disentangle content and style features via separate encoders.

Stochastic conditioning dropout training. During training, we unfreeze the Hierarchical Projector and fine-tune the DiT with LoRA as shown in Figure 3. With probability p we randomly drop either the style or the subject reference, forcing the model to solve pure subject-driven generation or pure style transfer tasks. This strategy preserves single-task capability while simultaneously exposing the network to a multi-task regime, enabling end-to-end learning of disentangled representations. The final multimodal input sequence z_2 is therefore expressed as:

$$z_2 = \text{Concatenate}(z_s, c, z_t, z_c), \tag{2}$$

We set $p = 0.25$ during training. Style tokens z_s are assigned the same positional indices as the text tokens c, while content tokens obtain their positions via UnoPE [37] using its diagonal layout. Consequently, USO can seamlessly handle both subject-driven and style-driven tasks on the proposed triplet dataset.

3.2.2. Auxiliary Style Reward

Although the above pipeline already formulates a unified customization model, one of our key insights is that learning to include desired features for one task helps the complementary task suppress those undesired features, thereby improving overall performance. To this end, we introduce a auxiliary style reward (ASR) for auxiliary style-driven task to boost style similarity and observe how it contributes to subject consistency. ASR alternates between computing a reward score and back-propagating the reward signal. Unlike traditional ReFL [41], which in text-to-image generation primarily considers text fidelity or visual appeal, ASR is tailored for the reference-to-image setting. It focuses on reinforcing the model to extract desired features from a reference image by directly computing a reward between the online outputs and the conditioning image. As shown in Figure 3, we define the reward score as the style similarity between the reference style image I^s_{ref} and the generated stylized image $\hat{I}_0$, measured by either a VLM-based filter or the CSD model $\mathcal{M}_{\text{RM}}(\cdot)$ [30, 40]. The reward loss is defined as:

$$\mathcal{L}_{\text{ASR}} = \mathbb{E}[\phi(\mathcal{M}_{\text{RM}}(I^s_{\text{ref}}, \hat{I}_0))] \tag{3}$$

where $\mathcal{Y} = \{y_i\}_{i=1}^n$ is the prompt set, ϕ maps reward scores to per-sample loss values, and $\hat{I}_0$ denotes the image gener-

Figure 5. Qualitative comparison with different methods on style-driven generation.

ated by the diffusion model with parameters θ corresponding to prompt y.

To mitigate potential reward hacking, we jointly optimize the model by including the original Flow-Matching training objective, which is computed as:

$$\mathcal{L}_{\text{Pre}} = \mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}[w(t)\|\boldsymbol{v}_\theta - \boldsymbol{v}_t\|^2] \tag{4}$$

where $w(t)$ is a weighting function, $\boldsymbol{v}_\theta$ denotes the neural network parameterized by θ, and the sampling process is from $t = T$ with $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to $t = 0$, by solving the PF-ODE via $d\boldsymbol{x}_t = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t)dt$. The final objective combines both losses:

$$\mathcal{L} = \mathcal{L}_{\text{Pre}} + \lambda\mathcal{L}_{\text{ASR}}, \quad \lambda = 0 \text{ before step } S,\ \lambda = 1 \text{ thereafter.} \tag{5}$$

As shown in Appendix Algorithm 1, we present the detailed ASR algorithm.

4. Experiments

4.1. Experiments Setting

USO unified benchmark. To enable a comprehensive evaluation, we introduce USO-Bench, a unified benchmark built from 50 content images (20 human-centric, 30 object-centric) paired with 50 style references. We further craft 30 subject-driven prompts that span pose variation, descriptive stylization, and instructive stylization, along with 30 style-driven prompts. We generate four images per prompt for both subject-driven and style-driven tasks, and a single image for the combined style-subject-driven task. This yields 6000 samples for subject-driven generation, 7040 for style-driven generation, and 29500 for the combined task; full construction details are provided in the supplementary material.

Evaluation metrics. For quantitative evaluation, we assess each task along three dimensions: (1) *subject consistency*, measured by the cosine similarity of CLIP-I and DINO embeddings following [37]; (2) *style similarity*, reported via the CSD score [30] for both style-driven and style-subject-driven generation, following [40]; and (3) *text–image alignment*, evaluated with CLIP-T for all three tasks.

Comparative methods. As a unified customization framework, USO is evaluated against both task-specific and unified baselines. For subject-driven generation, we benchmark RealCustom++ [20], RealGeneral [19], UNO [37], OmniGen2 [36], BAGEL [4], FLUX.1 Kontext dev [16], and Qwen-Image Edit [35]. For style-driven generation, we compare StyleStudio [17], DreamO [23], CSGO [40], InstantStyle [32], and DEADiff [26]. For the joint style-subject-driven setting with dual conditioning, we compare OmniStyle [33] and StyleID [3].

User study. We further conduct a comprehensive user study in Appendix Section 6.3.

4.2. Experimental Results

Subject-driven generation. As shown in Figure 4, the first two rows demonstrate that USO simultaneously satisfies both descriptive and instructive style edits while maintaining high subject consistency. In contrast, competing methods either fail to apply the style or lose the subject. The last two rows further illustrate USO's strength in preserving human appearance and identity; it adheres strictly to the textual prompt and almost perfectly retains facial and bod-

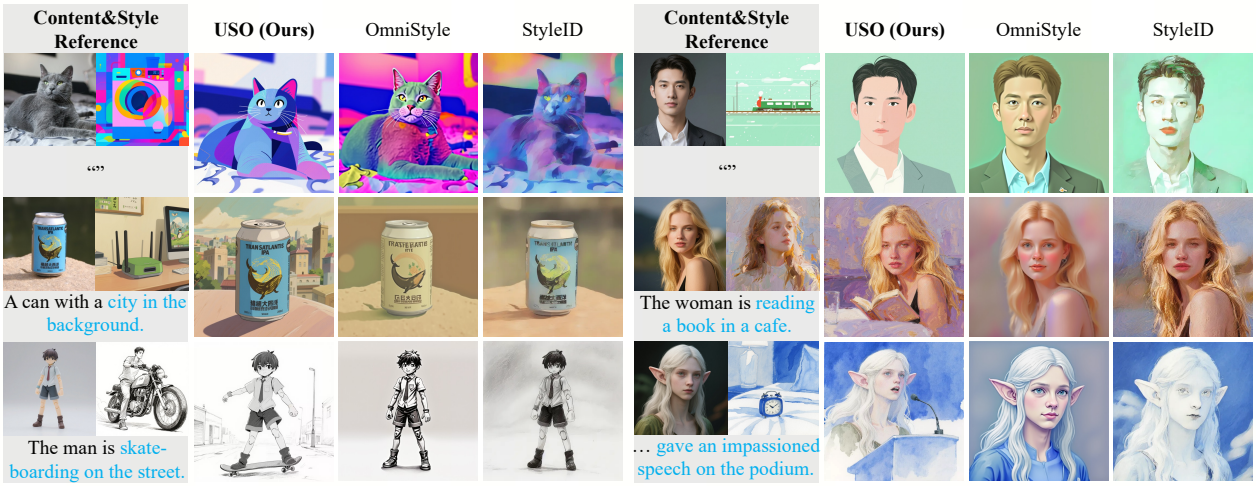


Figure 6. Qualitative comparison with different methods on style-subject-driven generation.

Method	Subject-driven generation			Style-driven generation		Style-subject-driven generation	
	CLIP-I↑	DINO↑	CLIP-T↑	CSD↑	CLIP-T↑	CSD↑	CLIP-T↑
RealCustom++ [12]	0.314	0.615	**0.303**	-	-	-	-
RealGeneral [19]	0.485	0.732	0.275	-	-	-	-
UNO [37]	0.605	0.789	0.264	-	-	-	-
BAGEL [4]	0.516	0.741	0.298	-	-	-	-
OmniGen2 [36]	0.475	0.723	0.302	-	-	-	-
FLUX.1 Kontext dev [16]	0.579	0.775	0.287	-	-	-	-
Qwen-Image Edit [35]	0.544	0.756	0.302	-	-	-	-
DreamO [23]	0.588	0.787	0.280	0.454	0.278	-	-
DEADiff [26]	-	-	-	0.462	0.274	-	-
InstantStyle-XL [32]	-	-	-	0.540	0.276	-	-
CSGO [40]	-	-	-	0.452	0.272	-	-
StyleStudio [17]	-	-	-	0.348	0.282	-	-
StyleID [3]	-	-	-	-	-	0.407	0.230
OmniStyle [33]	-	-	-	-	-	0.365	0.229
USO (Ours)	**0.647**	**0.804**	0.287	**0.556**	**0.286**	**0.492**	**0.283**

Table 1. Quantitative results on USO-Bench. We highlight the best and second-best values for each metric.

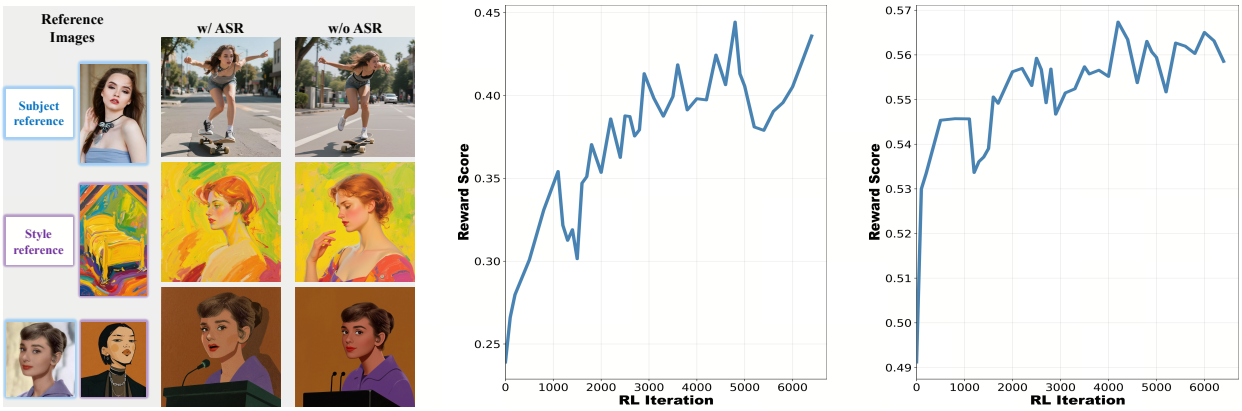
ily features, whereas other approaches fall short. When the prompt is “The man is reading a book in a cafe”, FLUX.1 Kontext dev [16] achieves decent facial similarity but carries copy-paste risks. As reported in Table 1, USO significantly outperforms prior work, achieving the highest DINO and CLIP-I scores and a leading CLIP-T score.

Style-driven generation. Figure 5 shows that USO outperforms task-specific baselines in preserving the original style, including global color palettes and painterly brushwork. In the last two rows, given highly abstract references such as material textures or Pixar-style renderings, USO handles them almost flawlessly while prior methods struggle, demonstrating the generalization power of our cross-task co-disentanglement. Quantitatively, Table 1 confirms that USO achieves the highest CSD and CLIP-T scores among all style-driven approaches.

Style-subject-driven generation. As illustrated in Figure 6, we evaluate USO on both layout-preserved and layout-shifted scenarios. When the input prompt is empty, USO not only preserves the original layout of the content reference but also delivers the strongest style adherence. In the last two rows, under a more complex prompt, USO simultaneously preserves the subject and identity consistency, matches the reference style, and aligns with the text, while other methods lag markedly and merely adhere to the text. Table 1 corroborates these observations, showing USO achieves the highest CSD and CLIP-T scores and substantially outperforms all baselines.

4.3. Ablation Study

Effect of auxiliary style reward (ASR). As shown in Figure 7(a), the middle column reveals a clear boost in style



(a) Qualitative comparison

(b) Reward on identity similarity (subject-driven)

(c) Reward on style similarity (style-driven)

Figure 7. Ablation study of ASR. ASR enhances identity consistency even though it relies solely on style reward.

Model	Subject-driven		Style-driven		Style-subject-driven	
	CLIP-I↑	CLIP-T↑	CSD↑	CLIP-T↑	CSD↑	CLIP-T↑
USO (Ours)	**0.647**	**0.287**	**0.556**	**0.286**	**0.492**	**0.283**
w/o ASR	0.619	0.283	0.491	0.281	0.413	0.280
w/o DE	0.594	0.269	0.491	0.280	0.382	0.277

Table 2. Ablation study of different components proposed in USO.

similarity for both style-driven and style-subject-driven tasks, with the identity of the woman and the painting style closely matching the reference images. Removing ASR leads to a sharp drop in the CSD score and simultaneous declines in CLIP-I and CLIP-T, as reported in Table 2. We further visualize the reward curves in Figure 7(b) and Figure 7(c); our method yields improvements in both identity and style similarity. Notably, we rely **solely on style reward signals** and introduce **no identity-specific supervision**; nevertheless, the unified model gains in identity consistency. By sharpening the model's ability to extract and retain desired features, ASR brings overall improvements across all tasks, validating our motivation.

Model	Subject-driven		Style-subject-driven	
	CLIP-I↑	CLIP-T↑	CSD↑	CLIP-T↑
USO (Ours)	**0.647**	**0.287**	**0.492**	**0.283**
UNO	0.605	0.264	-	-
UNO*	0.596	0.278	-	-
OmniStyle	-	-	0.365	0.229
OmniStyle*	-	-	0.382	0.277

Table 3. Quantitative results on USO-Bench. $*$ denotes models reproduced on our USO dataset.

Effect of disentangled encoder (DE). Replacing the disentangled encoders with a single shared VAE to encode both style and content images degrades nearly every metric (Table 2). We provide a qualitative comparison in Figure 10 of Section 6.3.3.

Effect of curated dataset. As shown in Table 3, we reproduce two representative task-specific methods, UNO [37] and OmniStyle [33], on our dataset to validate the effectiveness of the curated dataset. The reproduced OmniStyle even outperforms the original baseline, particularly in terms of CLIP-T, thanks to the layout-shifted triplets in the new dataset. Training UNO solely on the new dataset yields partial improvement, further confirming that both our method and the dataset contribute to the overall performance of USO.

5. Conclusion

In this paper, we present USO, a unified framework capable of subject-driven, style-driven, and joint style-subject-driven generation. We introduce a cross-task co-disentanglement paradigm that first constructs a systematic triplet-curation pipeline, then applies task disentanglement training on the curated triplets to formulate a unified customization model. Additionally, a auxiliary style reward learning paradigm is proposed to further boost performance. To comprehensively evaluate our method, we construct USO-Bench, which provides both task-specific and joint evaluation for existing approaches. Finally, extensive experiments demonstrate that USO sets new state-of-the-art results on subject-driven, style-driven, and their joint style-subject-driven tasks, exhibiting superior subject consistency, style fidelity, and text controllability.

Acknowledgment

This research is supported by National Natural Science Foundation of China under Grant 623B2094.

References

[1] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 13

[2] Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025. 3

[3] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. 6, 7

[4] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 6, 7

[5] 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 *ICML*, 2024. 3

[6] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024. 4

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 13

[8] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024. 4

[9] Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 5

[10] 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. 12

[11] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 3

[12] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: narrowing real text word for real-time open-domain text-to-image customization. In *CVPR*, pages 7476–7485, 2024. 2, 7, 13

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[14] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint arXiv:2402.12974*, 2024. 3

[15] Black Forest Labs. Flux: Official inference repository for flux.1 models, 2024. Accessed: 2025-02-07. 2, 3, 12

[16] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 3, 4, 6, 7

[17] Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, and Chi Zhang. Stylestudio: Text-driven style transfer with selective control of style elements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23443–23452, 2025. 2, 6, 7

[18] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023. 13

[19] Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. *arXiv preprint arXiv:2503.10406*, 2025. 6, 7

[20] Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom++: Representing images as real-word for real-time customization. *arXiv preprint arXiv:2408.09744*, 2024. 2, 6, 13

[21] Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom++: Representing images as real textual word for real-time customization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[22] Zhendong Mao, Mengqi Huang, Yijing Lin, Quan Wang, Lei Zhang, and Yongdong Zhang. Toward accurate image generation via dynamic generative image transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 2

[23] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025. 3, 6, 7

[24] 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 *ICLR*, 2024. 2, 3

[25] Senthil Purushwalkam, Akash Gokul, Shafiq Joty, and Nikhil Naik. Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. *arXiv preprint arXiv:2401.13974*, 2024. 13

[26] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Dead-

iff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8693–8702, 2024. 2, 3, 6, 7

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 3, 12, 13

[29] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 2

[30] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 5, 6

[31] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024. 3, 5, 13

[32] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 4, 6, 7

[33] Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7847–7856, 2025. 4, 6, 7, 8

[34] 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 *CVPR*, pages 15943–15953, 2023. 13

[35] 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. 6, 7

[36] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 4, 6, 7

[37] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 2, 3, 4, 5, 6, 7, 8, 12, 13

[38] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323*, 2021. 3

[39] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 13

[40] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 2, 4, 5, 6, 7

[41] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 5

[42] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[43] 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*, pages 11975–11986, 2023. 5, 12

[44] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*, pages 8069–8078, 2024. 5, 13