

The Consistency Critic: Correcting Inconsistencies in Generated Images via Reference-Guided Attentive Alignment

Ziheng Ouyang¹ Yiren Song² Yaoli Liu³ Shihao Zhu¹ Qibin Hou^{4,1*} Ming-Ming Cheng^{4,1} Mike Zheng Shou²

¹VCIP, Nankai University ²Show Lab, National University of Singapore

³State Key Laboratory of CAD&CG, Zhejiang University ⁴NKIARI, Shenzhen Futian

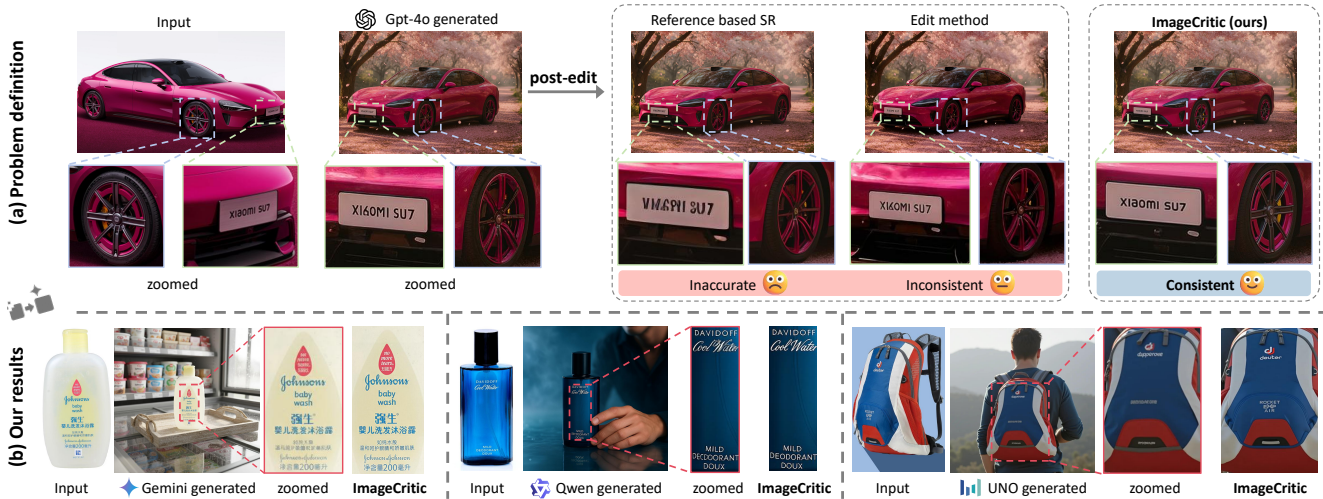


Figure 1. **Visual illustrations.** (a) illustrates that we first generate customized content using GPT-4o [1] and then apply different post-editing methods. Edit method [57] struggles to achieve fine-grained, consistent generation, while images processed by super-resolution methods [14] often exhibit noticeable detail inaccuracies. In contrast, our proposed ImageCritic corrects local details to ensure text and logo consistency while maintaining accurate spatial alignment, significantly improving the overall coherence of generated images. (b) We further apply our method to customized results generated by both state-of-the-art closed-source [10] and open-source models [57, 59]. After performing our correction, the fine-grained details of the generated images align precisely with those of the original objects, demonstrating the superior performance of our approach.

Abstract

Previous works have explored various customized generation tasks given a reference image, but they still struggle to generate consistent fine-grained details. In this paper, we aim to address the inconsistency problem in generated images by applying a reference-guided post-editing approach and present our ImageCritic. We first construct a dataset of reference-degraded-target triplets generated via VLM-based selection and explicit degradation, effectively simulating the common inaccuracies and inconsistencies observed in existing generation models. Furthermore, building on a thorough examination of the model’s attention mechanisms and intrinsic representations, we accordingly devise an attention alignment loss and a detail encoder to precisely rectify inconsisten-

cies. ImageCritic can be integrated into an agent framework to automatically detect inconsistencies and correct them through multi-round local editing in complex scenarios. Extensive experiments demonstrate that ImageCritic can effectively resolve detail-related issues in various customized generation scenarios, providing significant improvements over existing methods. Source code is available at: <https://ouyangziheng.github.io/ImageCritic-Page/>.

1. Introduction

In recent years, image generation systems [56] based on diffusion models have evolved from UNet architectures to Transformer-based Diffusion Transformer (DiT) models, with reference-guided generation becoming one of the key research trends, powering applications such as virtual try-

*Corresponding author.

on [15, 33, 35], image editing [8, 17, 21, 27, 52, 60–63, 63], and subject customization [4, 5, 17, 36, 42, 48, 59, 72, 73]. However, due to discrepancies introduced by VAE-based encoding and decoding in encoder–decoder frameworks, and the loss of shallow-layer information in decoder-only structures, existing models often produce inconsistent and inaccurate fine-grained details in the generated image relative to the reference image. As illustrated in Fig. 1, images generated using current state-of-the-art [1, 10, 57, 59] customized generation methods commonly exhibit inconsistencies or blurriness in textual and logo regions.

One naive approach to address this inconsistency between generated and reference images is to use reference-based super-resolution models, such as ReFIR [14], which refine blurry regions to improve alignment with the reference. However, as shown in Fig. 1(a), the inconsistencies in the generated images remain uncorrected, and in some local regions the errors become even more severe—falling short of the requirements for consistency-aware correcting. Another approach is to use multi-image editing models, such as Qwen-Image [57], to repair local details through text instructions. As shown in Fig. 1(a), the model fails to accurately locate and correct the target details. These demonstrate that, even with a reference image, enhancing the consistency of existing customized generation models remains a challenging problem.

We summarize the failure of the aforementioned models into two main issues. First, there is a lack of high-quality data that is capable of focusing on and improving fine-grained details. For example, datasets such as Subjects200K [48], UNO-1M [59], and X2I2 [58] mainly focus on maintaining the overall object consistency while neglecting subtle local details. To address this, we construct reference–target pairs that maintain strong consistency and simultaneously capture the inherent generation issues of current models. Specifically, we employ a VLM-based selection strategy to build reference–target pairs and further introduce Flux-Fill [23] to explicitly degrade local regions of the generated images, thereby simulating detail-level inconsistencies and completing our dataset construction.

In addition, existing models often fail to attend to, localize, and align fine-grained regions for precise correction. To tackle this problem, we visualize the attention maps of the reference and input images within the noise regions and observe strong coupling between them. To mitigate this, we introduce an attention alignment loss, which explicitly supervises the attention maps using region masks, enabling effective decoupling and automatic localization of the regions to be corrected. Furthermore, we analyze the intrinsic properties of our chosen base model. To improve generalization and reference-image understanding, we proposed a detail encoder, which explicitly embeds features for both reference and input images, fully leveraging the strengths

of the pre-trained editing model and further improving the consistency between generated outputs and inputs.

To improve usability, we further design an agent-based processing chain that automates the entire workflow, including consistency evaluation, discrepancy localization, reference retrieval, and image refinement. This enables multi-round human–agent interaction and a fully automated one-click refinement process, greatly enhancing the model’s ability to handle complex real-world scenarios.

As shown in Fig. 1, our method significantly improves the consistency of generated images. The refined results preserve the physical priors of relative spatial relationships between objects while maintaining environmental and lighting consistency. Meanwhile, local regions such as the wheel hub in Fig. 1(a) demonstrate a remarkable improvement in structural fidelity and fine-grained alignment.

Our contributions can be summarized as follows.

- To further enhance the consistency of existing generative models, we construct a reference-degraded-target dataset and establish a benchmark for complex object customization, providing valuable resources for high-consistency generation and editing research.
- To address the common issue of detail inconsistency and inaccuracy for existing customized generation models, we introduce an attention alignment loss for region-aware reference perception and a detail encoder for precise task understanding.
- We design an agent system that seamlessly integrates the critic model, enabling automated and intuitive multi-round editing while ensuring robust performance.

2. Related Work

Conditional image generation. Conditional image generation aims to improve controllability in diffusion models by introducing external conditions. Existing condition-guided diffusion models can be broadly categorized into two types of tasks: *condition-aligned generation* and *condition-referenced generation*. Condition-aligned generation focuses on abstract structural control, such as pose or shape alignment [30–32, 68]. In this setting, conditional inputs are strictly aligned with the output at the pixel level, achieving fine-grained spatial control. In contrast, condition-referenced generation emphasizes semantic-level and object-specific control rather than pixel-level correspondence. Representative works, such as IP-Adapter [65], Textual inversion [12], and others, introduce adapters or encoders [13, 22, 26, 41, 53, 69] to inject subject-related conditional information into the diffusion process, enabling reference subject consistency.

Unlike previous approaches that inject conditional signals directly at the feature level, DiT-based methods [3, 25, 44–46, 50, 51, 55, 58, 60, 61, 71] unify semantic and spatial

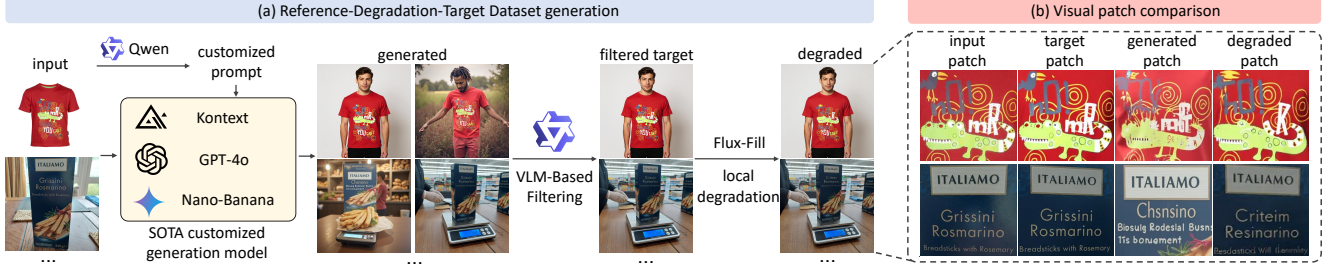


Figure 2. **Data curation pipeline.** (a) illustrates the complete pipeline of our approach, which involves generating customized images using existing state-of-the-art models, applying VLM-based filtering, and performing degradation. (b) shows local regions from our dataset, where the target patch aligns well with the input patch, and the degraded patch effectively simulates fine-grained inconsistencies in text and logo areas commonly seen between the input patch and the generated patch.

conditions into token sequences and integrate them with text tokens through multimodal attention and token concatenation mechanisms. To enhance subject consistency, DreamO [34] designs a router mechanism to focus attention on target subjects, and XVerse [4] adopts text-stream modulation to transform reference images into token-specific offsets. MOSAIC [42] further introduces fine-grained regional constraints via explicit semantic point correspondences. Despite these advances, inconsistency in fine-grained details remains a significant limitation.

LLM for Generation. Large Language Models (LLMs), such as ChatGPT [1] and Llama [49, 49], have demonstrated outstanding capabilities in the field of natural language processing. Meanwhile, Multimodal Large Language Models (MLLMs), including LLaVA [28], Claude, and GPT-4o, further integrate visual understanding, significantly enhancing the ability to process multimodal data. LLMs have now become a core component of vision-language tasks, where their powerful language understanding and reasoning abilities have driven advances in cross-modal semantic alignment and visual reasoning, as demonstrated in works such as VisProg [16] and ViperGPT [47]. At the same time, LLMs are widely adopted within agent frameworks [6, 29, 43, 64, 66], where they learn to invoke external tools to accomplish complex tasks such as visual interaction, speech processing [16], software development [38], and game operation [7]. With the deepening of multimodal training, the interconnections among generative agents have become increasingly integrated. LayoutGPT [9] can generate spatial layouts from textual prompts, GenArtist [54] and LLM Blueprints employ a “generate-then-edit” paradigm to optimize outputs, and LayerCraft [70] introduces an integrated multi-agent framework that unifies layout planning and object integration within a single system.

3. Data Curation

Under our problem setup, the input consists of the customized generated image to be repaired and a reference im-

age. Our goal is to enhance the fine-grained consistency of the generated image. However, given that existing models often suffer from detail misalignment issues, identifying more consistent input-output pairs to achieve a substantial performance improvement over current models remains a challenging problem.

We decompose the problem into two main components: constructing consistent reference-target pairs and generating images that reflect the inherent issues present in existing generative models. First, we employ a VLM-based filtering strategy to leverage the complementary strengths of both inputs, thereby constructing the most ideal ground-truth supervision. Then, we simulate common artifacts observed in current generative models, such as text rendering errors and logo mismatches, through controlled degradation.

As shown in Fig. 2(a), we begin by collecting a set of high-quality images and generating diverse prompts that include various objects, scenes, human-object interactions, and lighting conditions. Then, by using state-of-the-art text-to-image models, like Flux Kontext [24], GPT-4o [1], and Nano-Banana [10], we produce scene-rich images. These images are evaluated by Qwen-VL [2], which scores them in terms of visual quality and clarity. We retain only the top-performing samples. Next, Qwen is used again to annotate each image, followed by Grounding SAM [40] for object detection and segmentation. To mitigate potential mask errors from SAM, we further use Qwen to recheck and filter the outputs, ensuring consistency while preserving segmentation accuracy. This process yields a set of high-quality reference-target pairs. To simulate real-world degradation, we apply the Flux-Fill [23] model to actively corrupt the filtered high-quality images. Specifically, we randomly select subsets of the masked regions as inputs to Flux-Fill, prompting it to inpaint regions or alter elements, such as text or logos. The resulting degraded samples are then screened by Qwen to remove those with severe visual errors.

Through this data construction pipeline, we effectively address the two challenges described above and acquire 10k high-quality reference-degraded-target triplets. As il-

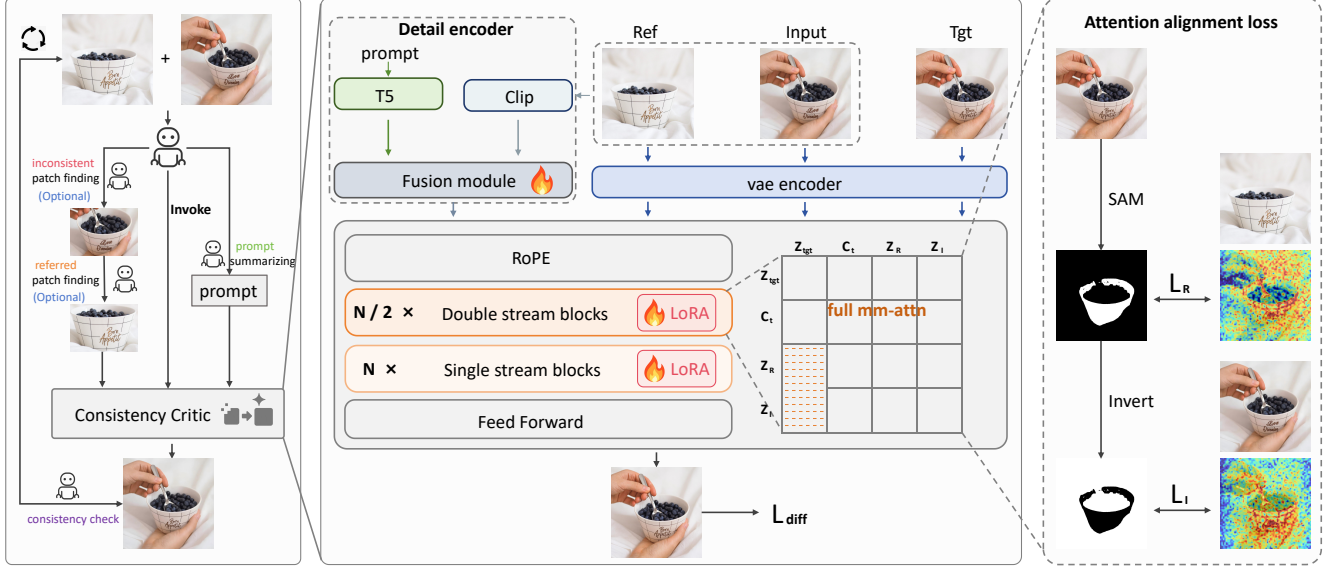


Figure 3. **Overview of the proposed ImageCritic.** We propose ImageCritic, which employs a Detail Encoder and an Attention Alignment Loss to enable the model to localize regions requiring restoration, thereby achieving high-quality and consistent image correcting. Furthermore, we develop a fully automated agent framework that supports both local patch restoration and multi-round correcting processes.

illustrated in Fig. 2(b), by comparing the input image patch and the target patch, one can observe that the filtered targets exhibit a high degree of consistency with the original input. Meanwhile, when comparing the model’s generated patch and the degraded patch, it can be seen that the degraded patches effectively reproduce common artifacts found in current generative models, including logo misalignment and text generation errors, which provide realistic supervision signals and a solid basis for subsequent model training.

4. Methodology

Our overall pipeline is illustrated in Fig. 3. As shown, the reference image and input image are fed into the detail encoder together with the prompt to obtain text tokens. These tokens are then concatenated with the image tokens encoded by the VAE and subsequently fed into the DiT for denoising. During training, we introduce an Attention Alignment Loss to ensure both the disentanglement and alignment of attentions between the condition inputs and the noisy target. During inference, we design an agent chain that, given a reference image and an input image to be corrected, automatically identifies the required patches and organizes the prompt. Both training and inference follow the same prompt format: “Use the object in IMG1 as a reference to be corrected, replace, or enhance the object in IMG2.” Here, “IMG1” and “IMG2” are trigger tokens corresponding to the reference image and the input image, respectively, while {object} denotes a simplified description of the target object identified from the reference image.

4.1. Preliminaries

The DiT architecture employs a multi-modal attention mechanism by concatenating text tokens $c_T \in \mathbb{R}^{M \times d}$ and noisy image tokens $z_{tgt} \in \mathbb{R}^{N \times d}$ as input, where d denotes the embedding dimension, and N and M are the numbers of image and text tokens, respectively. It projects the hidden states of these tokens into query Q , key K , and value V representations and perform attention:

$$\text{Attention}(Z) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d}} \right) V, \quad (1)$$

where $Z = [z_{tgt}, c_T]$ denotes the concatenation of the text tokens and the noisy image tokens. Recent models, such as the Flux Kontext image editing model [24], extend this formulation to $Z' = [z_{tgt}, c_T, z_c]$, which introduces an additional control token z_c to incorporate image-level conditioning. Here, z_c represents an image token corresponding to the image to be edited. We adopt Flux Kontext [24] as our base model to leverage its image control capability for fine-grained consistency correcting.

4.2. Consistency Critic

Attention alignment. Given the dataset obtained from Sec. 3, a straightforward idea is to fine-tune the base model using LoRA [20]. As shown in the first row of Fig. 4, after LoRA fine-tuning, the model is able to understand our intention and perform local correcting. However, for small text or fine-grained regions, the generated images still retain former elements from the input image, indicating that the overall consistency remains insufficient.

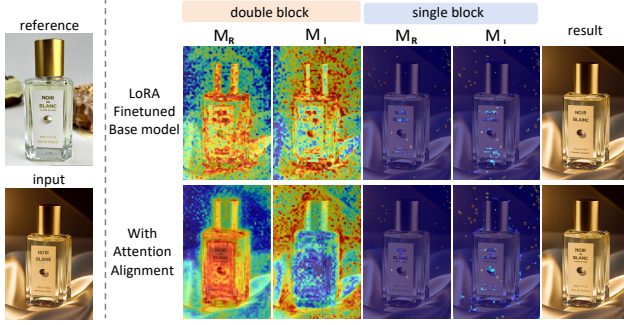


Figure 4. **Attention visualization.** We separately extract the noise attention maps with respect to the reference image and the input image to be corrected, denoted as M_R and M_I , respectively. The first row shows the results of the LoRA-finetuned base model, while the second row presents the results after applying the attention alignment loss. The first two columns correspond to the attention map of the double stream layer, and the last two columns correspond to the single stream layer. It can be observed that the attention alignment loss effectively promotes attention disentanglement.

As observed in [18, 19, 34], the image-to-image attention mechanism is capable of capturing spatial patterns during generation. Motivated by this, we further visualize the attention maps between the condition inputs and the generation target, as shown in Fig. 4. Specifically, the attention map is formulated as

$$M = \frac{Q_{c_i} K_{tgt}^T}{\sqrt{d}}, \quad i \in \{R, I\}, \quad (2)$$

where R and I denote the reference and input images, respectively, Q_{c_i} represents the condition tokens projected into the query space, and K_{tgt} denotes the noisy tokens projected into the key space.

By visualizing the influence of tokens from the reference and input branches on the noise latent, we observe that in our fine-tuned model, the attentions from these two branches are strongly coupled within the noise region. As a result, the input and reference tokens convey conflicting cues for local generation, preventing the model from accurately identifying the reference details necessary for precise correction. Consequently, the model either neglects modifications in these regions or performs inaccurate edits. To achieve the goal of consistency enhancement, we expect the input tokens to provide guidance on object-level details, while the reference tokens focus on background, lighting, and other global aspects. To realize this effect, we introduce an Attention Alignment Loss that enforces alignment at the attention level, explicitly guiding the model to query and reference the target regions appropriately.

Specifically, we first define a binary object mask B as:

$$B(p) = \begin{cases} 1, & p \in \text{background}, \\ 0, & p \in \text{subject}, \end{cases} \quad (3)$$

Subsequently, we incorporate an MSE-based alignment term into the DiT framework to optimize the attention distributions of both the reference and input branches with respect to the noise regions, defined as:

$$L_G = \frac{1}{n_l} \sum_{j=0}^{n_l-1} \|B \odot N(M_G^j)\|_2^2, \quad (4)$$

$$L_R = \frac{1}{n_l} \sum_{j=0}^{n_l-1} \|\bar{B} \odot N(M_R^j)\|_2^2, \quad (5)$$

where $\bar{B}$ denotes the complement of B , $\odot$ represents element-wise multiplication, j is the layer index, n_l is the total number of layers, and $N(\cdot)$ denotes a min-max normalization operator, which normalizes the input values to the range between zero and one. Finally, the overall training objective is formulated as:

$$L = L_{diff} + L_R + L_G. \quad (6)$$

Here, L_{diff} represents the rectified flow-matching loss originally adopted by the base model.

This design encourages the reference patches to focus primarily on the object regions rather than the background, while preventing the input patches from exerting excessive influence on the regions that require correction. Moreover, as observed in Fig. 4, in the double stream layers, the attentions from both the reference and input branches are concentrated and exhibit strong responses to the noise regions, while in the single block, the attention from the condition branch is more dispersed, lacking distinct regional features. Therefore, we apply the Attention Alignment Loss only at the double stream blocks of the model, and n_l is set to the number of double-stream block layers only.



Figure 5. **Effect of the detail encoder.** We find that when the input image exhibits structural differences from the reference image, the model fails to correctly identify the intended reference object, leading to inconsistent generation results.

Detail encoder. During training, we provide the model with two image inputs simultaneously and set the triggers “IMG1” and “IMG2” to represent each image, enabling precise prompt description. However, for the T5 text encoder,

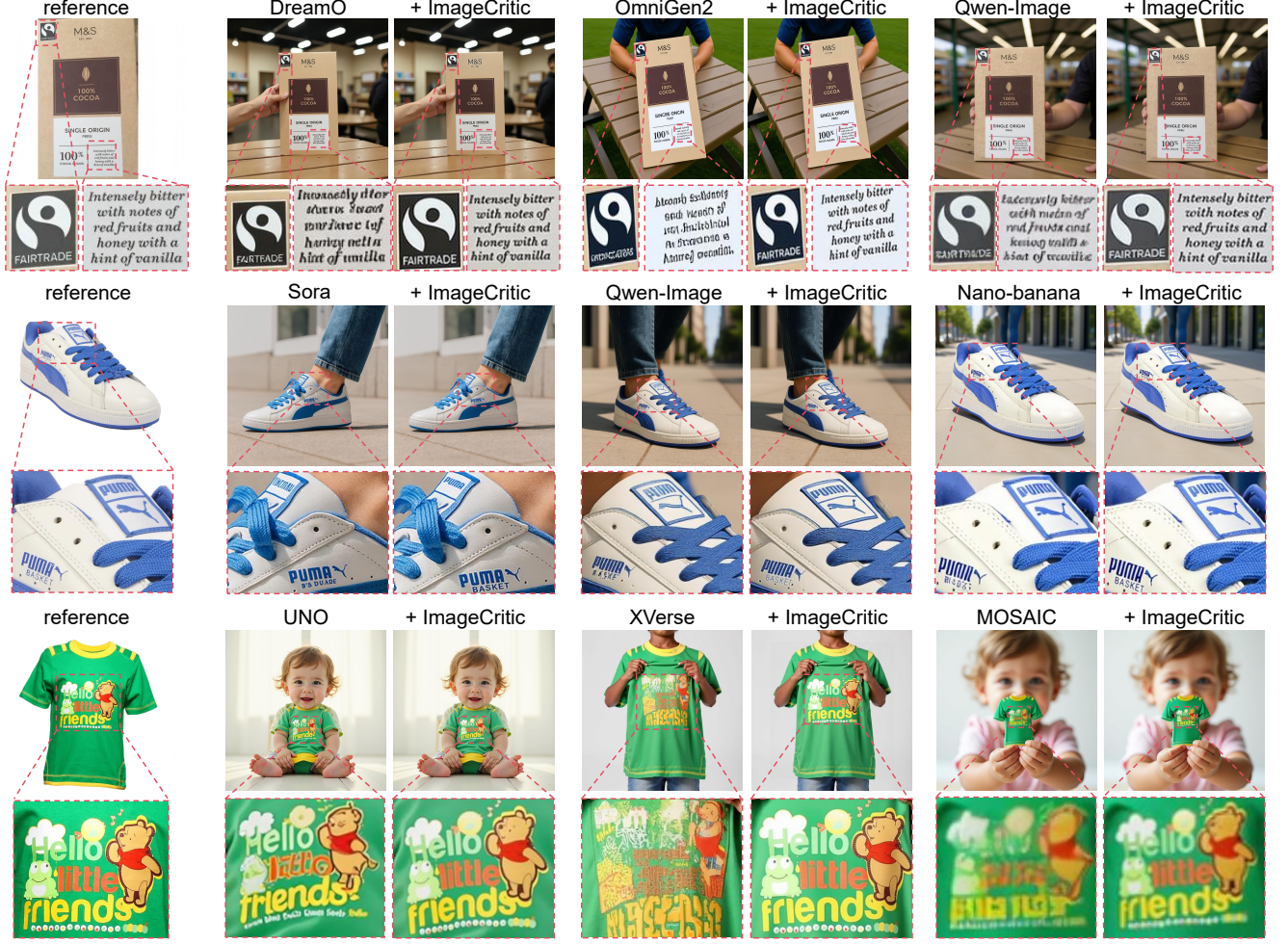


Figure 6. **Visual Results.** We present the results obtained by applying our method to the generated outputs of existing state-of-the-art customized generation models, including both open-source and closed-source variants. As shown, our method substantially enhances the consistency of the generated images.

the latent representation derived from the same trigger word remains identical across different inputs, which causes the model to struggle in associating different image inputs with their respective triggers, leading to incorrect outputs. As shown in Fig. 5, when the input image slightly differs in shape from the reference image, the model struggles to accurately identify the corresponding part of both condition images within the prompt, leading to erroneous generation.

Inspired by works such as PhotoMaker [26] and Dreamo [34], we design a Detail Encoder that couples the latent representation of the trigger word with that of the image, enhancing the model’s understanding of the reference image. This coupling mechanism allows the model to link textual triggers and visual content more effectively, thereby improving reference consistency in the generated results.

Specifically, let the prompt, after being processed by the T5 encoder, produce token representations $P \in \mathbb{R}^{M \times d_t}$, and let the flattened hidden representation of the image ob-

tained from CLIP [39] be $C_i \in \mathbb{R}^{1 \times d_c}$. From these tokens, we extract the hidden states $P_R, P_I \in \mathbb{R}^{1 \times d_t}$, corresponding to the trigger words “IMG1” and “IMG2”, which refer to the reference image and the input image to be corrected, respectively. Here, d_t and d_c denote the hidden dimensionalities of the T5 and CLIP models, respectively. We then construct a fused hidden representation:

$$P'_i = [P_i; C_i] \in \mathbb{R}^{1 \times (d_t + d_c)}, \quad i \in \{R, I\}. \quad (7)$$

Subsequently, P'_i is passed through a two-layer MLP and projected back to the original dimension, yielding $\tilde{P}_i$, which is then used to update the corresponding hidden states of the trigger words P_R and P_I in the prompt representation, respectively.

4.3. Agent Chain

In our experiments across various generative models, we observe that they tend to compress the reference image, re-

sulting in local regions where fine textual details become too small or even fail to render properly. However, in real-world scenarios, users often provide high-resolution reference images that are not fully utilized by such models. Moreover, existing models sometimes produce results in which the reference image occupies only a small portion of the generated output, which can negatively affect the correcting quality. To better preserve fine details, fully leverage high-resolution guidance, and enable more intuitive interaction, we develop an agent chain framework. Our system integrates multiple specialized agents through carefully designed prompts, with each agent responsible for specific subtasks—such as assessing content consistency, identifying inconsistent regions, selecting relevant reference patches, summarizing prompts, and performing final image correction (handled by the critic model). Specifically, we employ the Qwen-Agent [2] as a coordinator, which supervises the overall workflow, facilitates smooth communication between users and agents, and orchestrates cooperation among all components. While the system can operate autonomously, users can also interactively adjust input patches. For instance, if a user wishes to regenerate small textual details within a product region, the coordinator delegates this request to the corresponding agent for reprocessing, ensuring that the input patches align precisely with the user’s intent.

5. Experiments

5.1. Implementation details.

We adopt Flux.1-Kontext-dev [24] as our base model, upon which we conduct LoRA fine-tuning with a learning rate of 1×10^{-4} and a rank of 128, training on 2 GPUs with a batch size of 4 per GPU (total batch size = 8) for a total of 20,000 steps. After training, we enhance a subset of the generation targets with low consistency in the dataset using our proposed ImageCritic to further improve overall generation consistency. Details are provided in the Supplementary Material. The model was then retrained following the same settings described above.

5.2. Qualitative comparisons.

We employ both open-source models — XVerse [4], Dreamo [34], MOSAIC [42], OmniGen2 [58], UNO [59], and Qwen-Image [57] — as well as the closed-source models NanoBanana [10] and GPT-Image [1], to evaluate single subject customization performance. Subsequently, we apply our proposed method to enhance the consistency of the generated results. As illustrated in Fig. 6, existing approaches can generally preserve the primary subject; however, they often struggle to handle fine-grained local details, which limits their overall applicability. In contrast, our method achieves a significant improvement in consis-

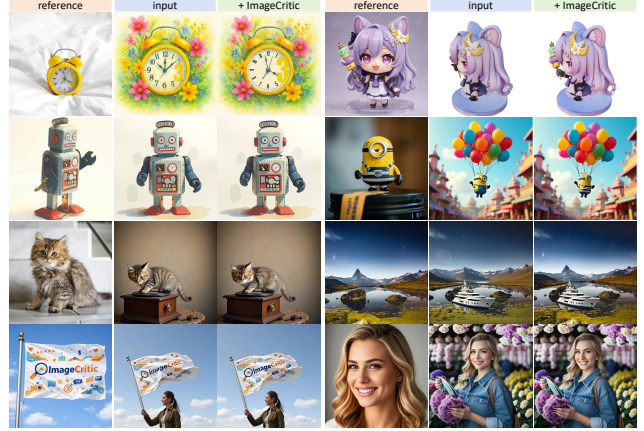


Figure 7. **Robust Generalization.** As shown in the figure, ImageCritic produces detail-rich and spatially coherent corrections across varying viewpoints and stylized settings, covering challenging cases such as non-rigid objects, faces, and backgrounds. Meanwhile, it preserves style and illumination consistency, rather than resorting to a naive copy-and-paste strategy.

tency while maintaining the original lighting, texture, and background fidelity of the generated images. In addition to these advantages, our approach further demonstrates strong robustness and generalization across diverse scenarios. As shown in Fig. 7, our method achieves highly consistent correcting across varying viewpoints, object categories, and diverse stylization settings, while faithfully preserving the background, illumination, and stylistic characteristics of the input images.

5.3. Quantitative comparisons.

Benchmark. We first evaluate our model on Dreambench++ [37]. Subsequently, to further examine the model’s performance on more challenging cases involving text, logos, and other complex visual details, we introduce a new benchmark, CriticBench, which specifically focuses on fine-grained detail preservation. Our benchmark consists of 300 images, including 200 intricate multilingual product images and 100 images of apparel and accessories. For each image, we utilize the Qwen-VL [2] model to generate editing instructions that most accurately align with the intended scene of the input image. Subsequently, we apply our method to these results and evaluate the quality of the final generated images.

Evaluation Metrics. We evaluate our approach on the aforementioned DreamBench++ [37] and CriticBench benchmarks, leveraging the methods introduced earlier for comparison. For each method, we compare each generated image with its corresponding reference image and compute metrics, including CLIP Image Score [39], DINO Score [67], and DreamSim [11]. Specifically, on our proposed CriticBench, for each method, we utilize Qwen-VL

Table 1. **Quantitative comparison on CriticBench.** Evaluation of consistency correcting across different generation models on our proposed CriticBench.

Method	CLIP-I $\uparrow$	DINO $\uparrow$	DreamSim $\downarrow$
Sora [1]	78.7 $\rightarrow$ 79.6 $+0.9$	68.4 $\rightarrow$ 69.2 $+0.8$	29.1 $\rightarrow$ 28.7 -0.4
Nano-Banana [10]	79.2 $\rightarrow$ 79.8 $+0.6$	66.5 $\rightarrow$ 66.9 $+0.4$	32.0 $\rightarrow$ 31.8 -0.2
Xverse [4]	76.5 $\rightarrow$ 79.9 $+3.4$	68.8 $\rightarrow$ 71.9 $+3.1$	34.3 $\rightarrow$ 31.4 -2.9
DreamO [34]	77.8 $\rightarrow$ 78.1 $+0.3$	67.7 $\rightarrow$ 68.2 $+0.5$	29.6 $\rightarrow$ 29.2 -0.4
MOSAIC [42]	74.6 $\rightarrow$ 77.1 $+2.5$	62.6 $\rightarrow$ 65.0 $+2.4$	35.2 $\rightarrow$ 31.4 -3.8
OmniGen2 [58]	78.8 $\rightarrow$ 79.3 $+0.5$	70.0 $\rightarrow$ 70.8 $+0.8$	27.7 $\rightarrow$ 27.0 -0.7
UNO [59]	77.6 $\rightarrow$ 78.9 $+1.3$	68.4 $\rightarrow$ 69.3 $+0.9$	33.6 $\rightarrow$ 32.1 -1.5
Qwen-Image [57]	77.9 $\rightarrow$ 78.2 $+0.3$	69.2 $\rightarrow$ 69.4 $+0.2$	30.3 $\rightarrow$ 30.1 -0.2

Table 2. **Quantitative comparison on DreamBench++.** Evaluation of consistency correcting across different generation models on DreamBench++.

Method	CLIP-I $\uparrow$	DINO $\uparrow$	DreamSim $\downarrow$
Xverse [4]	81.5 $\rightarrow$ 82.3 $+0.8$	64.5 $\rightarrow$ 66.3 $+1.8$	36.9 $\rightarrow$ 34.9 -2.0
DreamO [34]	83.3 $\rightarrow$ 83.4 $+0.1$	72.5 $\rightarrow$ 72.7 $+0.2$	27.2 $\rightarrow$ 27.1 -0.1
MOSAIC [42]	77.5 $\rightarrow$ 78.3 $+0.8$	57.6 $\rightarrow$ 59.2 $+1.6$	42.7 $\rightarrow$ 41.1 -1.6
OmniGen2 [58]	78.5 $\rightarrow$ 79.2 $+0.7$	62.0 $\rightarrow$ 64.5 $+2.5$	38.6 $\rightarrow$ 36.3 -2.3
UNO [59]	78.7 $\rightarrow$ 79.0 $+0.3$	62.7 $\rightarrow$ 63.2 $+0.5$	37.7 $\rightarrow$ 37.1 -0.6
Qwen-Image [57]	78.8 $\rightarrow$ 79.7 $+0.9$	61.8 $\rightarrow$ 64.9 $+3.1$	36.9 $\rightarrow$ 34.5 -2.4

Table 3. **Ablation study.** Quantitative evaluation of the contribution of each component in our proposed model.

AAL	DE	CLIP-I	DINO	DreamSim
$\times$	$\times$	77.9 $+0.3$	68.1 $+0.4$	31.3 -0.2
$\times$	$\checkmark$	78.3 $+0.7$	68.4 $+0.7$	30.6 -0.9
$\checkmark$	$\times$	78.3 $+0.7$	68.6 $+0.9$	30.6 -0.9
$\checkmark$	$\checkmark$	78.9 $+1.3$	68.9 $+1.2$	29.8 -1.7

[2] to perform semantic grounding on the corresponding objects within the generated images to enhance the evaluation of fine-grained detail consistency. The evaluation results on DreamBench++ [37] and CriticBench are presented in Tab. 1 and Tab. 2, respectively. As shown, our method substantially improves the consistency between the generated images and their corresponding reference images, demonstrating strong capability in recovering details.

5.4. Ablation analysis

We validate the effectiveness of our proposed dataset and conduct ablation studies on two key design components: Attention Alignment Loss (AAL) and Detail Encoder (DE).

Qualitative comparisons. As presented in Fig. 4 and Fig. 5, We compare the generation results of the model before and after adding our proposed module. As shown in Fig. 4, the results of the fine-tuned base model validate the effectiveness of our proposed dataset. However, simple fine-tuning still leads to deficiencies in subtle local details. After introducing the AAL, the attentions from the input and reference branches become concentrated on the required object and background regions, achieving attention disentanglement and significantly improving the model’s ability to correct local details. Meanwhile, as shown in Fig. 5, after incorporating the DE, the model’s capability

Table 4. **Agent chain discussions.** Comparison between our proposed agent chain and human annotations in locating required input and reference patches.

Metric	Mean IoU (%)	mAP@50 (%)
Score	75.3	88.4

to interpret complex inputs is enhanced, enabling it to accurately locate the target regions.

Quantitative evaluation, Following the setup in Sec. 5.3, we perform consistency correcting on the generated results of different methods using CriticBench, and report the average score improvements across all methods. As shown in Tab. 3, fine-tuning the base model on our dataset alone already improves the consistency of generated images, verifying the effectiveness of our proposed dataset. Moreover, introducing either component individually leads to additional performance gains, while combining both yields a further and more substantial improvement. These results confirm the effectiveness of our proposed modules and their synergistic effect.

5.5. Discussion

To evaluate the effectiveness of our proposed agent chain in locating input and reference patches, we manually annotated the target regions requiring correction along with their corresponding reference regions within the proposed benchmark. We then compared these human-defined bounding boxes with those automatically predicted by our agent, and quantified their alignment using the Intersection over Union (IoU):

$$\text{IoU} = \frac{|B_{\text{agent}} \cap B_{\text{human}}|}{|B_{\text{agent}} \cup B_{\text{human}}|}, \quad (8)$$

where B_{agent} and B_{human} represent the bounding boxes predicted by the agent and human annotators, respectively. Additionally, we also compute the mean Average Precision at a 50% IoU threshold (mAP@50), which assesses how accurately the agent detects and localizes regions whose overlap with the ground truth exceeds 50%. As shown in Tab. 4, our agent achieves a mean IoU of **75.3%** and a mean mAP@50 of **88.4%**, indicating its strong capability in accurately querying and localizing fine-grained details.

6. Conclusions

In this paper, we introduce ImageCritic, a novel post-editing correction framework designed to enhance the consistency of customized image generation, and construct a reference-degraded-target triplet dataset tailored to this task, enabling more accurate correcting. Extensive experimental results demonstrate that ImageCritic effectively mitigates inconsistencies in generated images, significantly improving detail consistency across both open-source and closed-source models.

Acknowledge. This work was supported by the National Natural Science Foundation of China (NSFC, 62225604), the Shenzhen Science and Technology Program (JCYJ20240813114237048), the “Science and Technology Yongjiang 2035” Key Technology Breakthrough Plan Project (2024Z120), and the Chinese Government-Guided Local Science and Technology Development Fund Projects (Scientific and Technological Achievement Transfer and Transformation Projects) (254Z0102G). Mike Shou does not receive any funding for this work.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 3, 7, 8
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 7, 8
- [3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Text to image generation with bidirectional multiway transformers. *Computational Visual Media*, 2025. 2
- [4] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025. 2, 3, 7, 8
- [5] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. 2
- [6] Yuming Chen, Jiangyan Feng, Haodong Zhang, Lijun Gong, Feng Zhu, Rui Zhao, Qibin Hou, Ming-Ming Cheng, and Yibing Song. Re-aligning language to visual objects with an agentic workflow. In *ICLR*, 2025. 3
- [7] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. 3
- [8] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 2
- [9] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023. 3
- [10] Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing gemini 2.5 flash image, our state-of-the-art image model. <https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/>, 2025. 1, 2, 3, 7, 8
- [11] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 7
- [12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2
- [13] Yan Gong, Yiren Song, Yicheng Li, Chenglin Li, and Yin Zhang. Relationadapter: Learning and transferring visual relation with diffusion transformers. *arXiv preprint arXiv:2506.02528*, 2025. 2
- [14] Hang Guo, Tao Dai, Zhihao Ouyang, Taolin Zhang, Yao-hua Zha, Bin Chen, and Shu-Tao Xia. Refir: Grounding large restoration models with retrieval augmentation. *arXiv preprint arXiv:2410.05601*, 2024. 1, 2
- [15] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. *arXiv preprint arXiv:2501.15891*, 2025. 2
- [16] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023. 3
- [17] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024. 2
- [18] Junjie He, Yifeng Geng, and Liefeng Bo. Unipor-trait: A unified framework for identity-preserving single- and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024. 5
- [19] Junjie He, Yuxiang Tuo, Binghui Chen, Chongyang Zhong, Yifeng Geng, and Liefeng Bo. Anystory: Towards unified single and multiple subject personalization in text-to-image generation. *arXiv preprint arXiv:2501.09503*, 2025. 5
- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4
- [21] Shijie Huang, Yiren Song, Yuxuan Zhang, Hailong Guo, Xueyin Wang, Mike Zheng Shou, and Jiaming Liu. Photodoodle: Learning artistic image editing from few-shot pairwise data. *arXiv preprint arXiv:2502.14397*, 2025. 2
- [22] 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*, pages 1931–1941, 2023. 2
- [23] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 2, 3
- [24] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 3, 4, 7
- [25] Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2671–2682, 2025. 2
- [26] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 2, 6
- [27] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. *arXiv preprint arXiv:2504.07960*, 2025. 2
- [28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3
- [29] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3
- [30] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2
- [31] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.
- [32] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 2
- [33] Ming Meng, Qi Dong, Jiajie Li, Zhe Zhu, Xingyu Wang, Zhaoxin Fan, Wei Zhao, and Wenjun Wu. HF-vton: High-fidelity virtual try-on via consistent geometric and semantic alignment. *arXiv preprint arXiv:2505.19638*, 2025. 2
- [34] 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, 5, 6, 7, 8
- [35] Haifeng Ni, Ming Xu, and Faming Fang. Itvton: Virtual try-on diffusion transformer based on integrated image and text. *arXiv preprint arXiv:2501.16757*, 2025. 2
- [36] Ziheng Ouyang, Zhen Li, and Qibin Hou. K-lora: Unlocking training-free fusion of any subject and style loras. *arXiv preprint arXiv:2502.18461*, 2025. 2
- [37] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 7, 8
- [38] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023. 3
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6, 7
- [40] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3
- [41] 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. 2
- [42] Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. *arXiv preprint arXiv:2509.01977*, 2025. 2, 3, 7, 8
- [43] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023. 3
- [44] Yiren Song, Danze Chen, and Mike Zheng Shou. Layer-tracer: Cognitive-aligned layered svg synthesis via diffusion transformer. *arXiv preprint arXiv:2502.01105*, 2025. 2
- [45] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572*, 2025.
- [46] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025. 2
- [47] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Pro-*

- ceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023. 3
- [48] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 2
- [49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3
- [50] Cong Wan, Xiangyang Luo, Zijian Cai, Yiren Song, Yunlong Zhao, Yifan Bai, Yuhang He, and Yihong Gong. Grid: Visual layout generation. *arXiv preprint arXiv:2412.10718*, 2024. 2
- [51] Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombe: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277*, 2025. 2
- [52] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024. 2
- [53] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 2
- [54] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395, 2024. 3
- [55] Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xi-angtai Li, and Yiren Song. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv:2505.21541*, 2025. 2
- [56] Yuxiang Wei, Yiheng Zheng, Yabo Zhang, Ming Liu, Zhilong Ji, Lei Zhang, and Wangmeng Zuo. Personalized image generation with deep generative models: A decade survey. *Computational Visual Media*, 2025. 1
- [57] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 1, 2, 7, 8
- [58] 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. 2, 7, 8
- [59] 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. 1, 2, 7, 8
- [60] Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28533–28543, 2025. 2
- [61] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 2
- [62] Ming Xie, Chenjie Cao, Yunuo Cai, Xiangyang Xue, Yugang Jiang, and Yanwei Fu. Anyrefill: A unified, data-efficient framework for left-prompt-guided vision tasks. *arXiv preprint arXiv:2502.11158*, 2025.
- [63] Zexuan Yan, Yue Ma, Chang Zou, Wenteng Chen, Qifeng Chen, and Linfeng Zhang. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. *arXiv preprint arXiv:2503.10270*, 2025. 2
- [64] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*, 2023. 3
- [65] Hu Ye, Jun Zhang, Sibao 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. 2
- [66] Xinbin Yuan, Jian Zhang, Kaixin Li, Zhuoxuan Cai, Lujian Yao, Jie Chen, Enguang Wang, Qibin Hou, Jinwei Chen, Peng-Tao Jiang, et al. Enhancing visual grounding for gui agents via self-evolutionary reinforcement learning. In *NeurIPS*, 2025. 3
- [67] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 7
- [68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2
- [69] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024. 2
- [70] Yuyao Zhang, Jinghao Li, and Yu-Wing Tai. Layercraft: Enhancing text-to-image generation with cot reasoning and layered object integration. *arXiv preprint arXiv:2504.00010*, 2025. 3
- [71] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025. 2
- [72] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. *arXiv preprint arXiv:2503.12885*, 2025. 2
- [73] Shihao Zhu, Bohan Cao, Ziheng Ouyang, Zhen Li, Peng-Tao Jiang, and Qibin Hou. Agebooth: Controllable facial

aging and rejuvenation via diffusion models. *arXiv preprint*
arXiv:2510.05715, 2025. [2](#)