UniCustom: Unified Visual Conditioning for Multi-Reference Image Generation

Yiyan Xu[1]*, **Qiulin Wang**[2]‡, **Wenjie Wang**[1]†, **Yunyao Mao**[2],
Xintao Wang[2], **Pengfei Wan**[2], **Kun Gai**[2], **Fuli Feng**[1]
[1]University of Science and Technology of China, [2]KlingAI Research
yiyanxu24@gmail.com, qiulin_wang@foxmail.com, wenjiewang96@gmail.com
https://yiyanxu.github.io/UniCustom/

Abstract

Multi-reference image generation aims to synthesize images from textual instructions while faithfully preserving subject identities from multiple reference images. Existing VLM-enhanced diffusion models commonly rely on decoupled visual conditioning: semantic ViT features are processed by the VLM for instruction understanding, whereas appearance-rich VAE features are injected later into the diffusion backbone. Despite its intuitive design, this separation makes it difficult for the model to associate each semantically grounded subject with visual details from the correct reference image. As a result, the model may recognize which subject is being referred to, but fail to preserve its identity and fine-grained appearance, leading to attribute leakage and cross-reference confusion in complex multi-reference settings. To address this issue, we propose ***UniCustom***, a unified visual conditioning framework that fuses ViT and VAE features *before* VLM encoding. This early fusion exposes the VLM to both semantic cues and appearance-rich details, enabling its hidden states to jointly encode the referred subject and corresponding visual appearance with only a lightweight linear fusion layer. To learn such unified representations, we adopt a two-stage training strategy: reconstruction-oriented pretraining that preserves reference-specific appearance details in the fused hidden states, followed by supervised finetuning on single- and multi-reference generation tasks. We further introduce a slot-wise binding regularization that encourages each image slot to preserve low-level details of its corresponding reference, thereby reducing cross-reference entanglement. Experiments on two multi-reference generation benchmarks demonstrate that UniCustom consistently improves subject consistency, instruction following, and compositional fidelity over strong baselines.

1 Introduction

Recent advances in text-to-image generation have substantially improved the fidelity, diversity, and controllability of visual synthesis [26, 11, 2, 19, 37, 20, 29, 30]. However, text alone is often insufficient for specifying fine-grained visual intent in practical creation scenarios. Users may wish to preserve and recombine concrete subjects from reference images, such as a particular person, object, garment, scene, or style. This has motivated increasing interest in *multi-reference image generation*, where a model is given multiple reference images and a textual instruction, and is expected to synthesize a coherent image that follows the instruction and preserves the specified subject identities and appearances from the references [43, 31, 3].

*This work was conducted during the author's internship at KlingAI.
†Corresponding authors.
‡Project Lead.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Multi-reference generation poses a unique challenge beyond standard text-to-image synthesis, where the textual instruction not only describes the target scene, but also specifies how subjects from different references should be selected, composed, and rendered [38, 10]. For example, an instruction may require "the woman from Picture 1" to wear "the hat from Picture 2" while interacting with "the dog from Picture 3". Successful generation therefore requires two essential capabilities: 1) ***Semantic grounding***, identifying which reference image, subject, or region is being referred to by each textual expression [3, 10]; 2) ***Visual binding***, associating each grounded subject with its corresponding appearance, identity, texture, and fine-grained attributes throughout the generation process [25]. These two capabilities are related but not equivalent. A model may correctly understand which subject is requested, yet still render that subject with attributes from another reference, leading to identity confusion, attribute leakage, missing entities, or incorrect compositions.

Recent VLM-enhanced diffusion models (*e.g.,* OmniGen2 [41], Qwen-Image-Edit [40], LongCat-Image-Edit [33]) provide a promising framework for this task by leveraging the multimodal understanding and instruction-following ability of Vision-Language Models (VLMs). As illustrated in Figure 1(a), a common design encodes reference images through two separate visual pathways. High-level ViT features are fed into the VLM together with the textual instruction, providing semantics for instruction understanding and semantic grounding. In parallel, VAE features, which preserve low-level visual details, are injected into the Diffusion Transformer (DiT) directly during generation to support faithful visual synthesis. This design is intuitive and has been widely adopted in reference-based generation [8, 13].

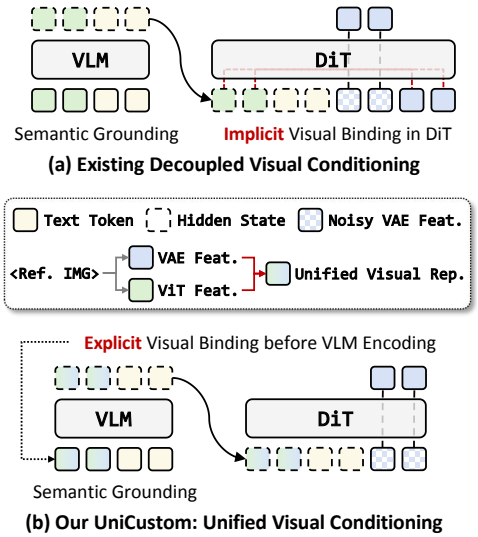

Figure 1: **Illustration of decoupled and unified visual conditioning.**

However, we argue that such decoupled design introduces a fundamental *grounding–binding gap*. Since the VLM only accesses semantic ViT features, its hidden states are well-suited for semantic grounding and instruction-level reasoning, but lack fine-grained appearance cues required for faithful rendering [8]. In contrast, VAE features, which preserve such appearance details, are injected only later into the DiT. Consequently, even when the VLM correctly grounds an instruction to the intended subjects, the DiT must still infer how these later-injected visual details should be associated with the VLM-encoded hidden states during generation, as illustrated in Figure 1(a). Such implicit visual binding becomes unreliable in multi-reference scenarios, where multiple subjects with similar appearances may coexist across references. The model may therefore follow the instruction at the semantic level while binding visual details to the wrong subject.

To address this issue, we propose ***UniCustom***, a unified visual conditioning framework for multi-reference image generation. As shown in Figure 1(b), UniCustom bridges the grounding–binding gap by fusing ViT and VAE features *before* VLM encoding. The resulting unified visual representation integrates ViT-derived semantic cues for reference grounding with VAE-derived appearance cues for faithful rendering. When jointly encoded with the textual instruction, it enables the VLM to produce hidden states that are both semantically addressable and appearance-aware, thereby providing the DiT with explicit semantic-visual correspondences during generation. Notably, our design is simple and lightweight: only a single linear fusion layer is sufficient to merge the two feature spaces, yielding substantial improvements in multi-reference grounding and visual consistency.

To make the unified visual representation effective for generation, we adopt a two-stage training strategy. In the pretraining stage, the model is optimized primarily on reconstruction-oriented tasks to establish semantic grounding, visual binding, and alignment between the unified visual representation and the DiT. During supervised fine-tuning, the model is mainly trained on single- and multi-reference image generation tasks, enabling the DiT to exploit VLM hidden states derived from the unified visual representation for reference-based synthesis. To further provide more structured conditioning signals for the DiT, we introduce slot-wise binding regularization on the VLM hidden states. This encourages each image slot to preserve reference-specific visual details while reducing cross-reference entanglement, allowing the DiT to parse and utilize multi-reference information more effectively.

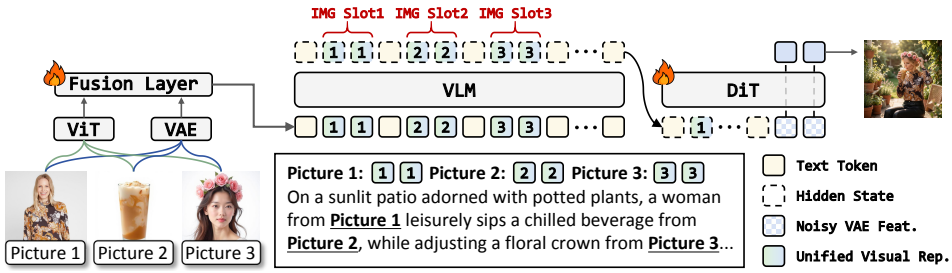

Figure 2: **Overview of UniCustom**. UniCustom fuses ViT and VAE features *before* VLM encoding, producing semantically addressable and appearance-aware hidden states for DiT generation.

To summarize, our contributions are as follows:

- We identify the *grounding–binding gap* in VLM-enhanced diffusion models for multi-reference image generation. In existing decoupled conditioning designs, the DiT must implicitly associate VLM-encoded subject semantics with separately injected appearance features, which becomes unreliable with multiple references.
- We propose ***UniCustom***, a unified visual conditioning framework that makes reference appearances semantically accessible. By fusing ViT and VAE features before VLM encoding, UniCustom produces hidden states that jointly encode the referred subject and its fine-grained visual details, thereby providing the DiT with more explicit semantic–appearance correspondences.
- We introduce a two-stage training strategy with slot-wise binding regularization to progressively learn reference-specific appearance preservation and adapt it to multi-reference generation. The reconstruction-oriented pretraining stage achieves a single-image reconstruction PSNR close to 30 dB, indicating that fused VLM hidden states can serve as an effective conduit for transmitting low-level details from VAE features to the DiT. Extensive experiments on two multi-reference image generation benchmarks further demonstrate that UniCustom outperforms existing methods.

2 Method

2.1 Model Architecture

As shown in Figure 2, our model builds upon VLM-enhanced diffusion models, in which the VLM encodes textual instructions and reference images into hidden states that serve as conditioning signals for the DiT during image generation. Our key architectural modification is a lightweight early-fusion module that injects VAE features into ViT features before VLM encoding. This design enables the VLM hidden states to incorporate both semantic grounding cues and fine-grained appearance details, thereby providing the DiT with more informative and reference-aware conditioning for generation.

Unified visual representation via early fusion. For each reference image, we extract a sequence of ViT features and VAE features, denoted as $\mathbf{F}^{\text{vit}} \in \mathbb{R}^{L \times d_{\text{vit}}}$ and $\mathbf{F}^{\text{vae}} \in \mathbb{R}^{L \times d_{\text{vae}}}$, respectively, where L represents the sequence length. Note that each reference image is resized so that the resulting ViT and VAE feature sequences have the same length and are spatially aligned. We concatenate the two feature sequences along the channel dimension and project them back to the ViT feature dimension using a lightweight linear fusion layer:

$$\mathbf{F}^{\text{uni}} = [\mathbf{F}^{\text{vit}}; \mathbf{F}^{\text{vae}}]\mathbf{W}_{\text{fuse}} + \mathbf{b}_{\text{fuse}}, \tag{1}$$

where $\mathbf{W}_{\text{fuse}} \in \mathbb{R}^{(d_{\text{vit}}+d_{\text{vae}}) \times d_{\text{vit}}}$ and $\mathbf{b}_{\text{fuse}} \in \mathbb{R}^{d_{\text{vit}}}$. The resulting unified representation $\mathbf{F}^{\text{uni}} \in \mathbb{R}^{L \times d_{\text{vit}}}$ has the same dimensionality as the original ViT features and can therefore be directly consumed by the pretrained VLM.

To preserve compatibility with the pretrained VLM at initialization, we adopt an identity-preserving initialization for the fusion layer. Specifically, the weights corresponding to the ViT feature dimensions are initialized as an identity mapping, while those corresponding to the VAE feature dimensions are initialized to zero:

$$\mathbf{W}_{\text{fuse}} = \begin{bmatrix} \mathbf{I}_{d_{\text{vit}}} \\ \mathbf{0}_{d_{\text{vae}} \times d_{\text{vit}}} \end{bmatrix}, \quad \mathbf{b}_{\text{fuse}} = \mathbf{0}_{d_{\text{vit}}}. \tag{2}$$

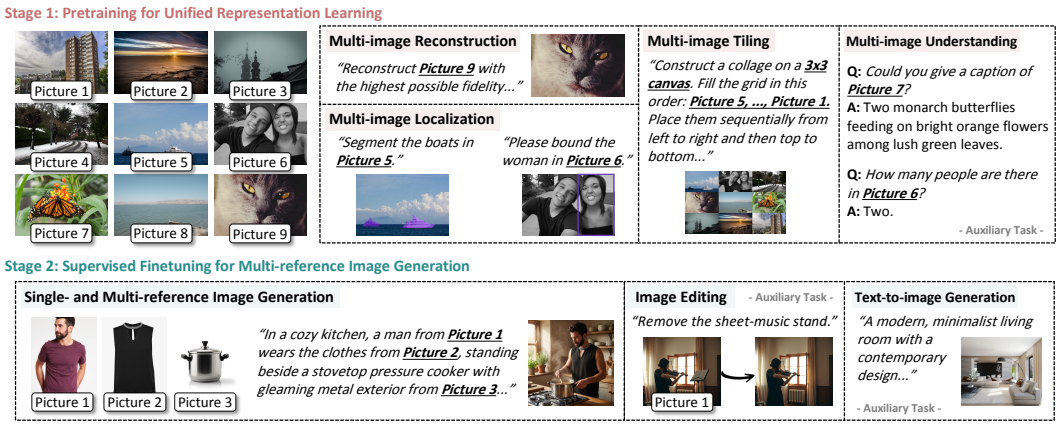

Figure 3: **Two-stage training strategy.** The first stage progressively learns a unified visual representation that supports fine-grained reference encoding, semantic grounding, and reliable textual-to-visual binding through reconstruction-oriented multi-image pretraining. The second stage further adapts the diffusion backbone to reference-based image generation, enabling instruction-following synthesis with single or multiple reference images while preserving the learned grounding and binding abilities.

Under this initialization, $\mathbf{F}^{\text{uni}}$ is initially equivalent to the original ViT feature. The model thus starts from the pretrained VLM-compatible visual feature space and gradually learns to incorporate fine-grained VAE appearance cues during training.

Multimodal encoding for DiT conditioning. Before VLM encoding, we organize the multimodal input sequence by interleaving explicit image identifiers with their corresponding unified visual representations. For N reference images, the input sequence is formatted as the following example:

> Picture 1: <$\mathbf{F}_1^{\text{uni}}$> Picture 2: <$\mathbf{F}_2^{\text{uni}}$> ... Picture N: <$\mathbf{F}_N^{\text{uni}}$> The woman from Picture 1 wears the hat from Picture 2... walking along a river with the dog from Picture N.

The explicit image identifiers act as language-level anchors for reference grounding, allowing textual expressions such as "the woman from Picture 1" or "the hat from Picture 2" to be associated with the corresponding visual representations. Since each visual representation has already integrated ViT semantics and VAE appearance cues, the VLM hidden states are both semantically addressable and appearance-aware, which serve as conditioning signals for the DiT, enabling the denoising process to rely on more explicit semantic-visual correspondences rather than inferring such bindings implicitly.

Positional encoding. For DiT conditioning, we adopt the Multimodal Rotary Position Embedding (MRoPE) from Qwen2.5-VL [1], which decomposes positional information into three components corresponding to the temporal, height, and width axes. For text positions, the same position index is assigned to all three components, making it equivalent to standard one-dimensional RoPE. For image positions, the temporal index is kept constant within each image, while the height and width indices are assigned according to the spatial location of each token.

2.2 Training Strategy

To effectively learn the proposed unified visual representation and align it with the diffusion denoising process, we adopt a two-stage training strategy. Throughout training, the VLM is kept frozen, and only the lightweight feature fusion layer and the DiT are optimized. This design preserves the pretrained multimodal understanding ability of the VLM, while allowing the model to progressively learn how to inject fine-grained visual details into the VLM hidden states and adapt the diffusion backbone to the resulting conditioning signals. Details about the training data can be found in Appendix B.

Stage 1: Pretraining for unified representation learning. In the first stage, we jointly optimize the fusion layer and the DiT with reconstruction-oriented tasks, aiming to learn unified visual representations that support reliable semantic grounding and visual binding, and adapt the DiT to

interpret the resulting VLM hidden states during denoising. As illustrated in Figure 3, the pretraining tasks include multi-image reconstruction, localization, and tiling. These reconstruction-oriented tasks require the model to identify the specified reference according to textual instructions and generate the corresponding visual output. In doing so, they encourage the VLM hidden states to encode reference semantics and fine-grained appearance cues, while simultaneously optimizing the DiT to parse these hidden states into the corresponding visual outputs.

To prevent the unified representation from being dominated by VAE-derived low-level visual details, we further mix a small proportion of multi-image understanding tasks into pretraining. Since the VLM is kept frozen throughout training, these tasks primarily regularize the fusion layer, encouraging the unified representation to remain compatible with the semantic structure expected by the VLM. After this stage, the model can transmit fine-grained appearance information from the VAE through the frozen VLM, maintain reliable textual-to-visual reference binding, and provide hidden states that are readily usable by the DiT denoising process.

Stage 2: Supervised finetuning for multi-reference image generation. In the second stage, we conduct Supervised Finetuning (SFT) on diverse single- and multi-reference image generation tasks, while freezing the fusion layer and updating only the DiT. This stage adapts the pretrained hidden states from reconstruction-oriented learning to realistic reference-based generation, enabling the DiT to synthesize instruction-following outputs that preserve the specified reference appearances. In addition, we mix in a small amount of text-to-image and image editing data to improve general instruction adherence, and retain a small portion of the pretraining tasks to mitigate forgetting of the reference grounding and binding ability acquired in the first stage.

Slot-wise Binding Regularization. When unified visual representations from multiple input images are jointly encoded with the text instruction, the resulting hidden states may contain entangled visual information across references (see empirical analysis in Section 3.4). This is partly due to the decoder-only formulation adopted by modern VLMs, where visual and textual tokens are processed under causal attention. Consequently, each token is contextualized by its preceding tokens. In multi-image inputs, the hidden states corresponding to a later image may therefore incorporate information from earlier references. While such contextualization is useful for integrating visual evidence with the textual instruction, it also blurs the localization of VAE-level visual details within their original positions, which makes the final hidden states less structured and harder for the DiT to parse.

To obtain more structured VLM hidden states, we introduce a slot-wise binding regularization during pretraining. As depicted in Figure 2, we define an image slot as the group of hidden states at the image-token positions assigned to a specific input image in the VLM sequence. Ideally, each slot should retain the VAE-level visual details of its corresponding image in a localized and decodable form. Specifically, for the i-th image, we take the hidden states $\mathbf{H}_i$ at its image slot and map them back to the VAE latent space using a single-layer projector $P(\cdot)$. The projected features are then supervised by the original VAE feature $\mathbf{F}_i^{\text{vae}}$ with a mean squared error loss:

$$\mathcal{L}_{\text{bind}} = \frac{1}{N} \sum_{i=1}^{N} \| P(\mathbf{H}_i) - \mathbf{F}_i^{\text{vae}} \|_2^2 . \tag{3}$$

This auxiliary objective encourages each image's visual details to remain recoverable from its own slot, making the final VLM hidden states more explicitly organized and easier for the DiT to parse. The projector is used only during pretraining and discarded afterwards.

3 Experiments

3.1 Implementation Details

We adopt Qwen2.5-VL [1] as the VLM backbone and initialize the DiT from LongCat-Image-Edit [33]. The VLM is kept frozen throughout training. Unless otherwise specified, all training samples are processed at resolutions no larger than 512×512. For evaluation, we report results on OmniContext [41] and MICo-Bench [39]. For our model and open-source baselines, we sample images at 512×512 resolution using the default inference configuration of each method, with a fixed random seed. For closed-source models, we use their native output resolution of 1024×1024, as they do not support direct generation at 512×512. More details can be found in Appendix A.

Table 1: **Quantitative comparison on OmniContext** [41], where "Char." and "Obj." denote "Character" and "Object", respectively. The best results in each group are highlighted in bold.

# OmniContext	SINGLE		MULTIPLE			SCENE			AVERAGE
	Character	Object	Character	Object	Char.+Obj.	Character	Object	Char.+Obj.	
Closed-source Models									
GPT-Image-2 [26]	**9.24**	**9.65**	**9.13**	**9.50**	**8.94**	**9.27**	**9.25**	**8.95**	**9.24**
Nano Banana 2 [11]	8.98	9.44	8.85	9.26	8.81	8.80	8.59	8.67	8.92
Open-source Models									
FLUX-Kontext [dev] [17]	7.09	7.35	2.49	5.46	4.46	2.86	3.89	3.40	4.62
UNO [43]	7.51	7.99	4.56	7.41	6.53	4.00	5.76	5.77	6.19
USO [42]	7.71	7.68	2.91	7.27	5.61	3.88	6.49	6.04	5.95
BAGEL [7]	7.12	7.70	5.86	7.07	6.79	4.98	6.05	6.03	6.45
OmniGen2 [41]	8.04	7.87	7.22	7.76	7.51	7.52	7.30	7.57	7.60
Qwen-Image-Edit-2509 [40]	8.11	**9.01**	**7.85**	**8.53**	7.60	5.88	7.38	7.12	7.69
LongCat-Image-Edit [33]	**8.24**	8.63	6.69	8.07	6.99	5.80	6.73	6.98	7.26
UniCustom (ours)	8.06	7.51	7.78	7.99	**7.86**	**7.86**	**7.72**	**7.92**	**7.84**



Figure 4: **Qualitative comparison on OmniContext [41].**

3.2 Main Results

Quantitative evaluation. Overall, UniCustom achieves the best performance among open-source models on both OmniContext [41] and MICo-Bench [39]. As shown in Tables 1 and 2, the gains are especially pronounced in multi-reference, scene-level, Object, HOI, and De&Re settings, which require accurate reference grounding, subject-level appearance preservation, and compositional reasoning across multiple inputs. These results verify that fusing ViT and VAE features before VLM encoding, combined with the proposed two-stage training strategy, is effective for improving semantic grounding and visual binding in multi-image reference tasks. As a result, UniCustom can better model complex relationships among referenced subjects and maintain subject consistency under diverse generation scenarios. Despite the remaining gap to closed-source models, UniCustom sets a strong open-source baseline on both benchmarks, demonstrating the promise of the proposed unified visual conditioning.

Table 2: **Quantitative comparison on MICo-Bench** [39], where "HOI" and "De&Re" denote "Human-object Interaction" and "Decomposition & Recomposition", respectively. The best results in each group are highlighted in bold.

# MICo-Bench	Object	Person	HOI	De&Re	Overall
Closed-source Models					
GPT-Image-2 [26]	66.93	60.56	59.02	62.69	61.78
Nano Banana 2 [11]	**68.46**	**65.88**	**64.88**	**70.28**	**67.42**
Open-source Models					
FLUX-Kontext [dev] [17]	21.40	14.33	12.67	7.24	12.51
UNO [43]	42.20	15.15	27.99	41.93	32.43
USO [42]	38.18	15.57	20.25	35.91	27.37
BAGEL [7]	39.23	16.77	20.99	46.75	31.62
Qwen-Image-Edit-2509 [40]	37.33	**28.01**	22.06	10.53	21.67
LongCat-Image-Edit [33]	40.55	19.40	24.47	14.86	22.78
UniCustom (ours)	**54.30**	18.12	**40.51**	**50.29**	**41.71**

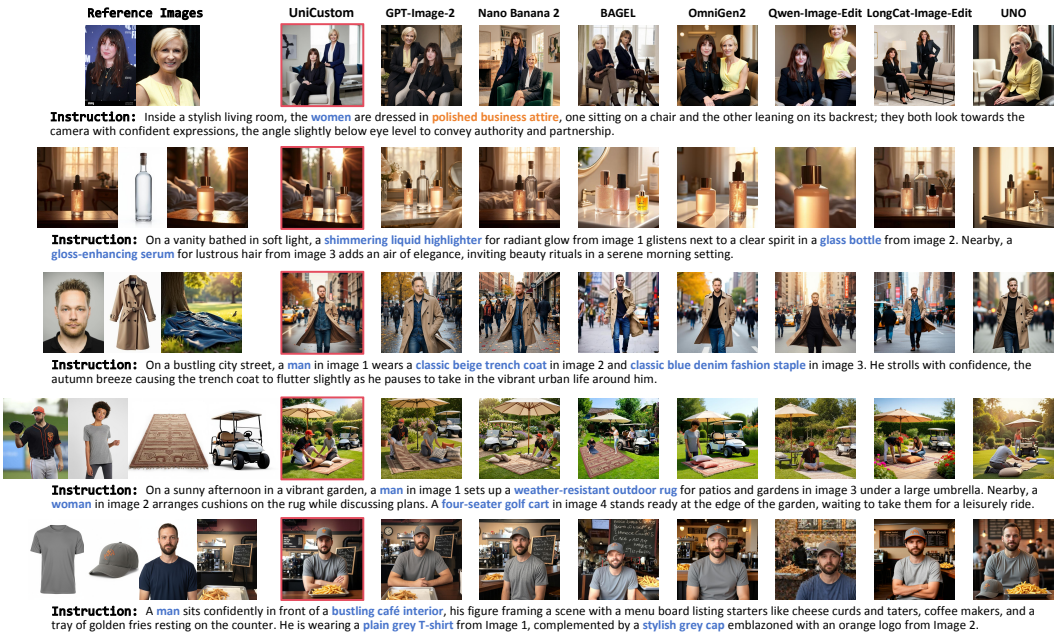

Figure 5: **Qualitative comparison on MICo-Bench [39].**

Qualitative evaluation. We provide qualitative comparisons with competitive baselines on OmniContext and MICo-Bench in Figure 4 and Figure 5, covering challenging multi-reference scenarios involving humans, objects, clothing, scenes, and spatial relationships. Compared with existing baselines, UniCustom better preserves the visual identities and fine-grained attributes from the reference images while following complex textual instructions to model diverse relationships among multiple references, including subject interactions, spatial arrangements, role assignments, and object-attribute correspondences. For example, UniCustom can faithfully generate interactions such as hugging or facing each other, spatial layouts such as one subject sitting while another leans on the chair backrest, and multi-reference scenes involving people and objects with coherent spatial arrangements. These results demonstrate the superiority of UniCustom in complex multi-reference generation, where it better avoids subject missing and attribute confusion while achieving stronger instruction following than competing methods. More cases can be found in Figure 11 and Figure 12 in the Appendix.

3.3 Reference Grounding and Binding Analysis

To further analyze how UniCustom leverages multiple visual references, we visualize the internal attention maps of the DiT in multi-reference image generation. As shown in Figure 6, given references containing a woman, a striped shirt, and a creamy blended beverage, UniCustom accurately attends to the corresponding regions in different reference images when generating each target component, demonstrating the effectiveness of UniCustom in semantic grounding and visual binding across multiple reference images. We also present the corresponding VLM response, which correctly identifies the woman, the striped short-sleeved shirt, and the creamy blended beverage, indicating that the fused VAE and ViT features are well understood by the VLM and support accurate instruction parsing. These results show that UniCustom can reliably associate textual expressions with their visual references and generate coherent outputs following complex multi-image instructions.

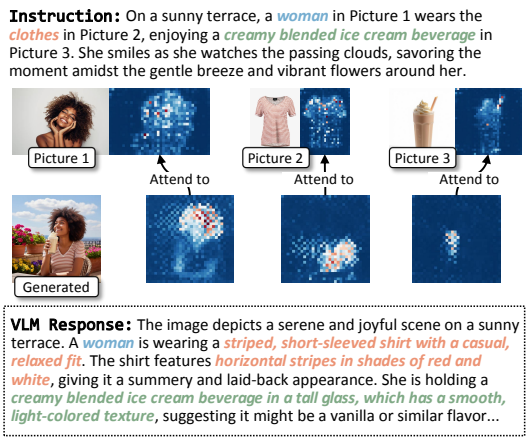

Figure 6: **Attention visualization of UniCustom.**

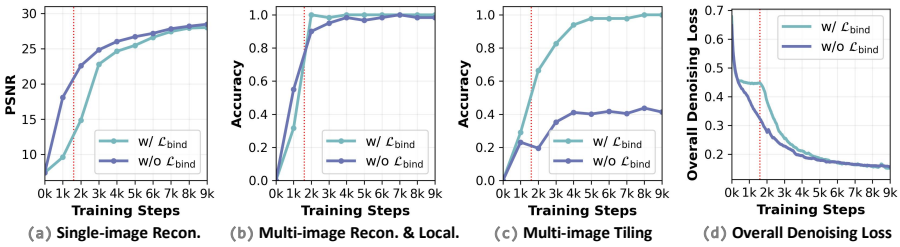

Figure 7: **Effect of slot-wise binding regularization**, where "Recon." and "Local." denote "Reconstruction" and "Localization", respectively.

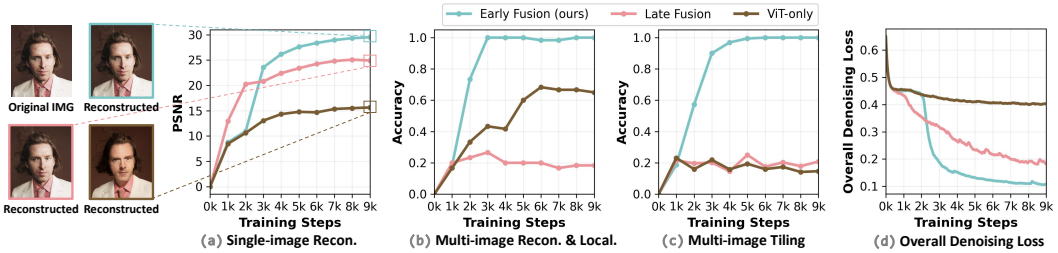

Figure 8: **Effect of different fusion strategies**, where "Recon." and "Local." denote "Reconstruction" and "Localization", respectively.

3.4 Ablation Study

We conduct ablations in the pretraining stage to validate two key design choices: slot-wise binding regularization and early fusion. We report single-image reconstruction Signal-to-Noise Ratio (PSNR) [12], and multi-image reconstruction, localization, and tiling accuracies, where the accuracies evaluate only image selection and placement rather than pixel-level reconstruction fidelity. These metrics measure whether the model can preserve visual details, bind them to parseable hidden states, and correctly use multiple input images, which are essential for multi-reference image generation; failures at pretraining will directly limit downstream generation quality.

Effect of slot-wise binding regularization. Figure 7 reveals distinct learning dynamics with and without slot-wise binding regularization during pretraining. Without $\mathcal{L}_{\text{bind}}$, the model quickly learns to pass through VAE-level visual details, leading to faster early improvement in single-image reconstruction as shown in Figure 7(a). However, this early visual-detail transmission does not translate into effective multi-image indexing, as reflected by slow learning in multi-image selection and consistently poor tiling accuracy in Figure 7(b) and (c). We attribute this behavior to the entangled multi-image hidden states produced by decoder-only VLMs. In models such as Qwen2.5-VL, the causal attention mechanism allows each token position to aggregate information from preceding tokens. When multiple images are serialized into a single sequence, visual details from different images can become mixed in the resulting hidden states, rather than remaining image-wise separable. Since these hidden states serve as the conditioning signals for the DiT, such entanglement makes it difficult to extract the relevant visual details needed for generation.

Slot-wise binding regularization alleviates this issue by encouraging localized image-slot binding, thereby producing more structured and parseable hidden states for resolving multi-image inputs. With $\mathcal{L}_{\text{bind}}$, each image slot is explicitly regularized to bind to the visual details of its corresponding image. Therefore, the model first learns a structured slot-image correspondence before substantial reconstruction improvement emerges. This explains why the generation loss changes only marginally in the early stage, especially before 2K steps, as shown in Figure 7(d). Once this indexing structure is established, the DiT can parse the VLM hidden states more effectively, leading to a sharp loss decrease and clear gains in multi-image reconstruction, localization, and tiling accuracy.

Effect of different fusion strategies. We compare our early fusion strategy with two variants during pretraining: late fusion, which injects VAE features into the corresponding image slots after VLM encoding, and ViT-only, which excludes VAE features entirely. As shown in Figure 8(a), late fusion substantially improves single-image reconstruction over ViT-only and converges faster

in the early training stage, indicating that directly injected VAE features can effectively provide fine-grained appearance information to the DiT. However, this improvement in reconstruction fidelity does not translate into reliable multi-image selection. Surprisingly, late fusion performs even worse than ViT-only on multi-image reconstruction and localization, as shown in Figure 8(b), while both variants exhibit consistently low tiling accuracy in Figure 8(c). Although late fusion explicitly assigns appearance-rich VAE features to their corresponding image slots, these features do not participate in the VLM's contextualization of visual references and textual instructions. Consequently, the encoded VLM hidden states are formed without access to fine-grained appearance information, leaving the DiT to coordinate separately encoded semantic and appearance cues. This separation may hinder effective reference grounding and visual information utilization. In contrast, early fusion achieves both the highest reconstruction quality and nearly perfect multi-image reconstruction, localization, and tiling accuracy. Together, these results suggest that effective multi-reference generation requires not only access to fine-grained appearance details, but also their integration into the VLM's contextualized representations, highlighting the importance of unified visual conditioning for bridging the grounding–binding gap.

4 Related Work

Multi-reference Image Generation. To enable multi-reference image generation, early methods mainly rely on per-subject optimization [36] or adapter-based feature injection [46, 18]. More recently, in-context conditioning has emerged as a more flexible paradigm. OminiControl [32] shows that DiT can inherently encode visual references, motivating methods such as UNO [43]. Subsequent works further improve this paradigm through attention constraints, as in MOSAIC [31] and DreamO [25]; modulation-based control, as in TokenVerse [10] and XVerse [3]; and reinforcement learning, as in UMO [5] and PSR [35]. Building on this trend, VLM-enhanced methods such as OmniGen2 [41], Qwen-Image-Edit [40], and Canvas-to-Image [6] leverage VLMs to improve instruction understanding and reference-aware generation. UniCustom further advances VLM-enhanced conditioning by unifying semantic ViT features and appearance-rich VAE features within the VLM conditioning stream, rather than processing them through decoupled pathways, yielding unified conditioning signals that jointly capture subject-level cues and reference-specific visual details for faithful generation.

VLM-enhanced Diffusion Models for Image-to-Image Generation. VLM-enhanced diffusion models have recently advanced image-to-image generation by incorporating multimodal understanding into diffusion-based synthesis. Some methods use VLMs to provide auxiliary guidance for diffusion models, such as input image encoding in BLIP-Diffusion [18], multimodal prompt understanding in Kosmos-G [27], and instruction refinement in MGIE [9]. Others move toward unified multimodal generation and editing frameworks, including Step1X-Edit [24], OmniGen2 [41], Qwen-Image-Edit [40], and LongCat-Image-Edit [33], where instruction understanding and image-to-image synthesis are integrated in a single system. While these methods differ in architecture and scope, they commonly rely on separate pathways for semantic visual features and appearance-rich generative features. This design leaves the correspondence between textual instructions and low-level visual details to be inferred implicitly during generation. UniCustom differs by explicitly aligning semantic and generative visual information before VLM encoding, yielding representations that are both semantically grounded and appearance-aware.

5 Conclusion

We presented ***UniCustom***, a unified visual conditioning framework for multi-reference image generation that addresses the grounding–binding gap in existing VLM-enhanced diffusion models. By fusing semantic ViT features and appearance-rich VAE features before VLM encoding, UniCustom enables the resulting hidden states to jointly encode instruction-level subject cues and reference-specific visual details. Combined with a two-stage training strategy and slot-wise binding regularization, this design improves subject consistency, instruction following, and compositional fidelity in complex multi-reference scenarios. Experiments on OmniContext and MICo-Bench demonstrate that UniCustom achieves strong performance among open-source methods. More broadly, our results show that appearance-rich VAE information can be effectively propagated through a frozen VLM and leveraged by the diffusion backbone for faithful generation. This suggests a promising direction for designing unified visual representations that connect multimodal understanding and generation.

Acknowledgements and Disclosures

This work was supported by the National Natural Science Foundation of China (U25A20340, U25B2071). This work was conducted during the first author's internship at KlingAI, which provided computational resources for the research. The authors declare no other competing interests.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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.

[2] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

[3] Bowen Chen, Haomiao Sun, Li Chen, Xu Wang, Daniel Du, Xinglong Wu, et al. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *Advances in Neural Information Processing Systems*, 38:7437–7459, 2026.

[4] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[5] 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.

[6] Yusuf Dalva, Guocheng Gordon Qian, Maya Goldenberg, Tsai-Shien Chen, Kfir Aberman, Sergey Tulyakov, Pinar Yanardag, and Kuan-Chieh Jackson Wang. Canvas-to-image: Compositional image generation with multimodal controls. *arXiv preprint arXiv:2511.21691*, 2025.

[7] 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.

[8] Yufan Deng, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Angtian Wang, Shenghai Yuan, Yiding Yang, Bo Liu, Haibin Huang, and Chongyang Ma. Cinema: Coherent multi-subject video generation via mllm-based guidance. *arXiv preprint arXiv:2503.10391*, 2025.

[9] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)*, 44(4):1–11, 2025.

[11] Google. Nano banana 2: Combining pro capabilities with lightning-fast speed. `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`, 2026.

[12] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[13] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

[14] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–629, 2023.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[16] Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. Nohumansrequired: Autonomous high-quality image editing triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6059–6068, 2026.

[17] 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.

[18] Dongxu Li, Junnan Li, and Steven Hoi. BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[19] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *International Conference on Learning Representations*, volume 2026, pages 86729–86758, 2026.

[20] Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Yanbin Hao, James Ma, Fuli Feng, et al. Speed: Scalable, precise, and efficient concept erasure for diffusion models. In *International Conference on Learning Representations*, volume 2026, pages 66777–66803, 2026.

[21] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[22] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[25] 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. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

[26] OpenAI. Introducing chatgpt images 2.0. `https://openai.com/index/introducing-chatgpt-images-2-0/`, 2026.

[27] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[28] Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. *arXiv preprint arXiv:2510.19808*, 2025.

[29] Junxiang Qiu, Lin Liu, Shuo Wang, Jinda Lu, Kezhou Chen, and Yanbin Hao. Accelerating diffusion transformer via gradient-optimized cache. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17608–17617. IEEE, 2025.

[30] Junxiang Qiu, Shuo Wang, Jinda Lu, Lin Liu, Houcheng Jiang, Xingyu Zhu, and Yanbin Hao. Accelerating diffusion transformer via error-optimized cache. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9588–9597, 2025.

[31] 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. In *The Fourteenth International Conference on Learning Representations*, 2026.

[32] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14940–14950. IEEE, 2025.

[33] Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

[34] Unsplash. The unsplash dataset. `https://github.com/unsplash/datasets`, 2025.

[35] Shulei Wang, Longhui Wei, Xin He, Jianbo Ouyang, Hui Lu, Zhou Zhao, and Qi Tian. Psr: Scaling multi-subject personalized image generation with pairwise subject-consistency rewards. *arXiv preprint arXiv:2512.01236*, 2025.

[36] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28759–28768. IEEE, 2025.

[38] Yuran Wang, Bohan Zeng, Chengzhuo Tong, Wenxuan Liu, Yang Shi, Xiaochen Ma, Hao Liang, Yuanxing Zhang, and Wentao Zhang. Scone: Bridging composition and distinction in subject-driven image generation via unified understanding-generation modeling. *arXiv preprint arXiv:2512.12675*, 2025.

[39] Xinyu Wei, Kangrui Cen, Hongyang Wei, Zhen Guo, Bairui Li, Zeqing Wang, Jinrui Zhang, and Lei Zhang. Mico-150k: A comprehensive dataset advancing multi-image composition. *arXiv preprint arXiv:2512.07348*, 2025.

[40] 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.

[41] 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.

[42] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025.

[43] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18682–18692, 2025.

[44] Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang YU, Xingjun Ma, and Yu-Gang Jiang. Withanyone: Toward controllable and ID consistent image generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[45] Qinhong Yang, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Lu Yuan, Gang Hua, and Nenghai Yu. Hq-50k: A large-scale, high-quality dataset for image restoration. *arXiv preprint arXiv:2306.05390*, 2023.

[46] 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.

[47] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

[48] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

[49] Yi Zhang, Bolin Ni, Xin-Sheng Chen, Heng-Rui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi-Min Hu. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. *arXiv preprint arXiv:2510.13795*, 2025.

A Implementation Details

During pretraining of UniCustom, we jointly optimize the fusion layer and the DiT with a learning rate of 5×10^{-5} for 18K steps. The training batches are uniformly sampled from four task groups: multi-image reconstruction, localization, tiling, and understanding. For supervised finetuning, we freeze the fusion layer and only update the DiT. We train for another 18K steps with a learning rate of 1×10^{-5}. The finetuning mixture consists of 10% pretraining tasks, 5% text-to-image generation, 10% image editing, 25% single-reference generation, and 50% multi-reference generation. This mixture preserves the reference grounding ability acquired during pretraining while adapting the model to open-ended reference-based image generation. The model is trained using 128 GPUs.

Note that OmniGen2 [41] is omitted from the quantitative comparison on MICo-Bench in Table 2 since it does not support generation conditioned on more than five reference images.

B Training Data

Multi-image reconstruction. We first curate a large-scale image pool from open-source datasets, including HQ-50K [45], MultiID-2M [44], HumanArt [14], FFHQ [15], and Unsplash [34]. Based on this corpus, we construct multi-image reconstruction data for pretraining. Each sample consists of multiple candidate images paired with an instruction that specifies the target image to be reconstructed (*e.g.,* "*Reconstruct Picture i with the highest possible fidelity...*"). To avoid resolution-based shortcuts, we partition images into resolution buckets and sample candidates from the same bucket, which encourages the model to ground the target image from the specified image identifier, rather than exploiting superficial resolution cues.

Multi-image localization. We construct multi-image localization data from COCO2017 [23]. Starting from single-image bounding-box and segmentation samples, we extend them into multi-image localization tasks following the same protocol as multi-image reconstruction. Each sample contains multiple candidate images with the same resolution, and the instruction specifies the target image to be localized (*e.g.,* "*Please bound the woman in Picture i*", "*Segment the boats in Picture i.*"), which strengthens semantic binding while preserving localization ability. By requiring localization under an explicitly specified reference index, this task promotes semantic grounding between the instruction and the target image, while reinforcing visual binding between the selected reference and its localized content.

Multi-image tiling. Based on the image pool used for multi-image reconstruction, we further construct multi-image tiling data. Specifically, we sample multiple images from the same resolution bucket and arrange them into grid-based collages with layouts of 2×2, 2×3, 3×2, and 3×3. The instruction specifies both the canvas layout and the ordering of input images (*e.g.,* "*Construct a collage on a 2×3 canvas. Fill the grid using images in this order: Picture 5,..., Picture 1. Place*

This survey evaluates images generated by AI for a multi-reference image generation task.

The survey contains 60 questions. In each question, you will see:
1. A text prompt describing the requested image;
2. One or more reference images showing the subjects or visual elements that should appear in the result;
3. Three AI-generated images labeled A, B, and C.
Please select the generated image that best follows the text prompt while preserving the identity and visual appearance of the subjects in the reference images.

 # Evaluation Criteria
Please consider the following criteria:
1. Prompt Following: How well the generated image follows the requested content, actions, relationships, composition, and setting.
2. Subject Consistency: How accurately the subjects preserve the identity and visual appearance shown in the reference images.
3. Image Quality: Whether the image is coherent, natural-looking, detailed, and free of visible artifacts.
4. Aesthetics: The overall visual appeal of the image, including its composition, lighting, color, and style.

Please give greater importance to Subject Consistency and Prompt Following when making your choice. Select one answer only for each question. There are no right or wrong answers — we are interested in your visual judgment. Please complete the survey using a desktop or laptop computer so that you can examine the images carefully.

Figure 9: **Instructions for human evaluation.**

them sequentially from left to right and then top to bottom..."). This task requires the model to follow explicit ordering constraints across multiple images, thereby strengthening semantic grounding to the image identifier and visual binding between each image and its assigned spatial position.

Multi-image understanding (auxiliary task). We construct multi-image understanding data from the captioning and general VQA subsets of Honey-Data-15M [49]. Following the same multi-image augmentation protocol, we convert single-image samples into multi-image tasks by adding candidate images and rewriting the instruction to query a specified reference (*e.g.,* "*Answer the question based on Picture i...*") This promotes semantic grounding to the designated image while retaining image-level understanding ability in multi-reference scenarios.

Single- and multi-reference generation. We collect reference-based generation data from Echo-4o-Image [47], Nano-Consistent-150K [47], MICo-150K [39], and an internal dataset. After filtering and reformatting, these data further adapt the model from reconstruction- and understanding-oriented pretraining to reference-based generation, where the model must ground textual instructions to the specified references and preserve the corresponding visual identities during synthesis.

Text-to-image generation and image editing (auxiliary task). Following UniWorld-V1 [22], we include a small amount of text-to-image data from BLIP-3o [4] and Open-Sora Plan [21], with image editing data from Pico-Banana-400K [28], ImgEdit [48], and NHR-Edit [16]. These samples serve as auxiliary SFT data that broaden the instruction distribution beyond reference-based prompts, while the training objective remains focused on single- and multi-reference image generation.

C Human Evaluation

We conducted a human evaluation on 60 cases, including 30 from OmniContext and 30 from MICo-Bench. We recruited 50 participants through CloudResearch [1] to evaluate all 60 cases, yielding 3,000 judgments in total. For each case, outputs from UniCustom, Qwen-Image-Edit, and LongCat-Image-Edit were presented in randomized order. Participants selected the best one by jointly considering prompt following, subject consistency, overall visual quality, and aesthetics. The full text of instructions and the evaluation interface are presented in Figure 9 and Figure 10. The resulting preference rates with 95%

Table 3: **Human evaluation results.**

Qwen-Image-Edit	LongCat-Image-Edit	UniCustom
$0.291^{\pm 0.015}$	$0.142^{\pm 0.018}$	$\mathbf{0.567}^{\pm 0.022}$

[1] https://www.cloudresearch.com/.

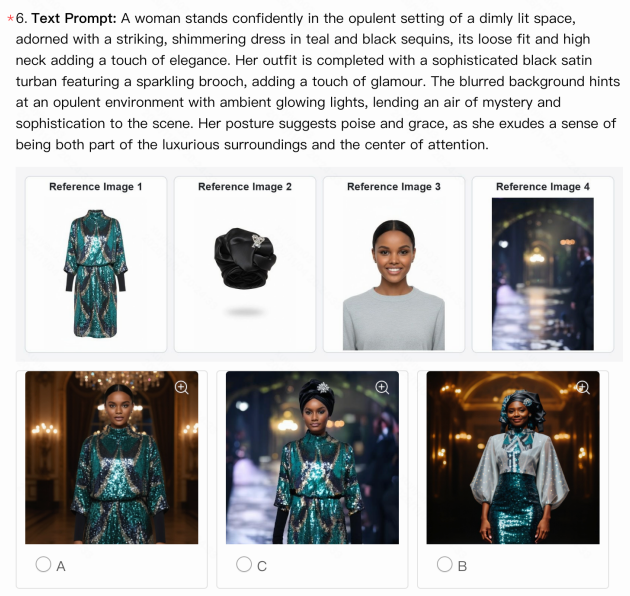

Figure 10: **Demonstration of human evaluation interface.**

confidence intervals are presented in Table 3. UniCustom received the highest overall preference, which is consistent with the quantitative results.

D Feature Propagation Analysis

Although UniCustom achieves high reconstruction fidelity through unified visual conditioning, it remains unclear how fine-grained VAE information is preserved and propagated through the frozen VLM. To better understand this process, we analyze the representation alignment and information recoverability of image-slot hidden states across different VLM layers.

Representation alignment. We first compute linear Centered Kernel Alignment (CKA) between intermediate VLM hidden states and the original ViT and VAE features on 200 held-out images. As a control, we also evaluate ViT-only features processed by the same frozen VLM. As presented in Table 4, the fused representations initially exhibit high alignment with ViT features, consistent with our identity-preserving initialization. Interestingly, their alignment with VAE features progressively increases from 0.541 before VLM encoding to 0.938 at the final layer, whereas ViT-only representations maintain consistently low alignment. This suggests that the learned fusion layer enables VAE information to be effectively propagated through the frozen VLM.

Table 4: **Layer-wise CKA analysis.**

	Pre-VLM	**Layer 7**	**Layer 14**	**Layer 21**	**Post-VLM**
Fused → VAE	**0.541**	**0.556**	**0.684**	**0.738**	**0.938**
ViT-only → VAE	0.363	0.364	0.369	0.358	0.351
Fused → ViT	**0.912**	**0.892**	**0.836**	**0.828**	**0.540**
ViT-only → ViT	1.000	0.995	0.983	0.972	0.940

Representation recoverability. Since CKA measures representational similarity rather than information recoverability, we further train layer-wise linear probes to predict VAE latents from intermediate hidden states, which are then decoded into images using the frozen VAE decoder. We report the reconstruction PSNR on 500 held-out images. As shown in Table 5, the fused representations achieve a PSNR of 29.00 dB, substantially outperforming ViT-only representations and approaching the VAE reconstruction upper bound (33.26 dB). These results indicate that fine-grained appearance

information remains linearly recoverable after VLM encoding, supporting the effectiveness of unified visual conditioning in preserving appearance-aware information for DiT generation.

Table 5: **Layer-wise linear probing analysis.**

	Pre-VLM	Layer 7	Layer 14	Layer 21	Post-VLM
Fused	**27.86**	**27.50**	**27.65**	**27.74**	**29.00**
ViT-only	12.45	12.47	12.35	12.09	11.87

E Qualitative Results

Multi-reference image generation. We provide more generated examples of UniCustom on OmniContext [41] and MICo-Bench [39] in Figure 11 and Figure 12. These examples further validate the effectiveness of UniCustom in multi-reference image generation, showing that its unified visual conditioning helps preserve fine-grained reference appearances while maintaining accurate semantic grounding and coherent subject composition.

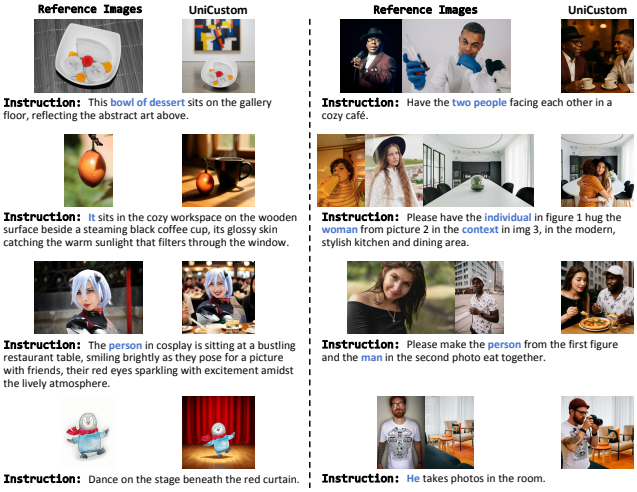

Figure 11: **More generated examples of UniCustom on OmniContext [41].**

Image localization. As illustrated in Figure 7, the single-image reconstruction PSNR demonstrates that low-level details encoded in the VAE features can be effectively transmitted through the VLM hidden states. In addition, the nearly perfect accuracy on multi-image reconstruction, localization, and tiling suggests that UniCustom establishes reliable semantic grounding and visual binding across multiple references. Specifically, the DiT can effectively parse the hidden states, follow the given instruction, and select or recombine the corresponding reference images as required. In Figure 13, we further present qualitative results for image localization. These examples show that UniCustom can accurately follow instructions to localize the specified subject, which is a crucial capability for downstream multi-image reference tasks.

Image editing and text-to-image generation. We further present qualitative results of UniCustom on image editing and text-to-image generation in Figures 14 and 15, respectively. Although UniCustom incorporates only a small amount of image editing and text-to-image data as auxiliary tasks to improve instruction adherence, it demonstrates strong generalization across both settings. For image editing, UniCustom performs well on a diverse range of tasks, including object addition, object replacement, background modification, color editing, object removal, and style transfer. For text-to-image generation, UniCustom also produces visually plausible and semantically aligned results, indicating that its learned capabilities extend beyond multi-reference generation.

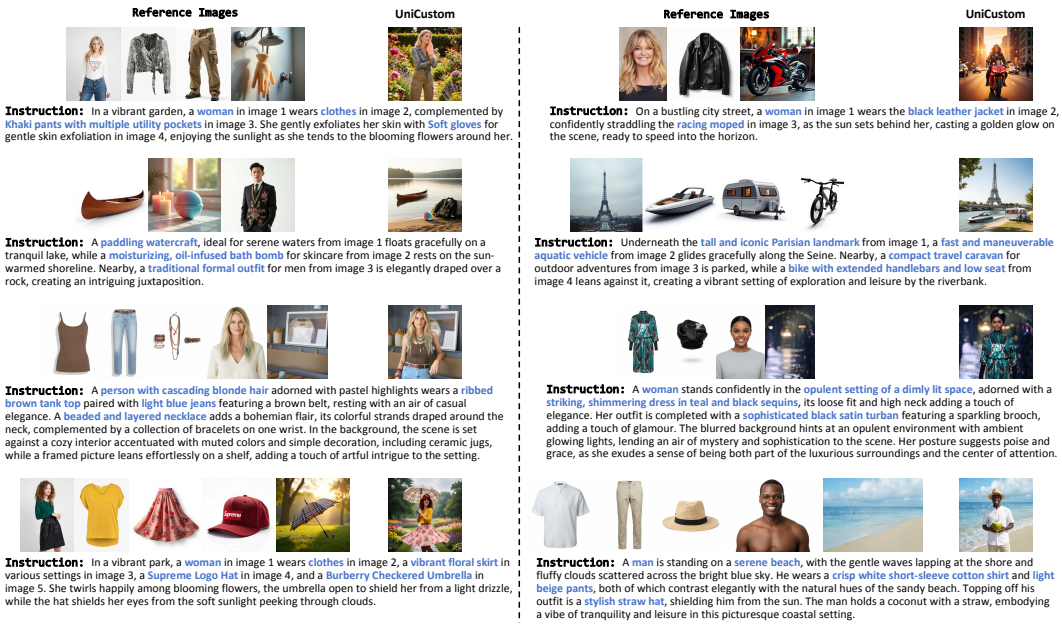

Figure 12: **More generated examples of UniCustom on MICo-Bench [39].**

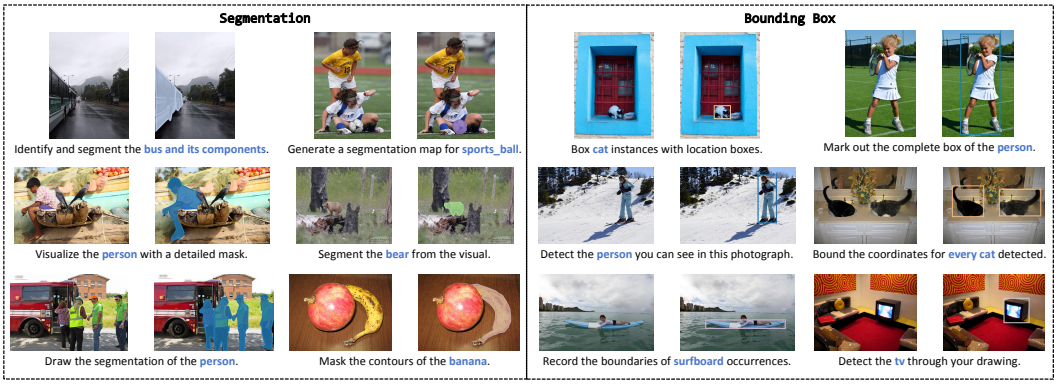

Figure 13: **Generated examples of UniCustom on image localization.**

F Limitations and Future Work

UniCustom achieves strong multi-reference image generation performance among open-source models, demonstrating the promise of unified visual conditioning for reference-based generation. Nevertheless, a gap remains compared with closed-source systems, particularly in highly realistic human identity preservation, complex object interactions, and fine-grained scene consistency. We regard this gap not as an inherent limitation of the framework, but as evidence of several important directions for future improvement.

First, UniCustom currently employs a lightweight fusion module that integrates spatially aligned ViT and VAE features through a simple linear layer. This design offers an initial exploration of combining understanding-oriented and generation-oriented visual representations within a unified conditioning pathway. Our results indicate that feature fusion before VLM encoding is effective; however, the current fusion mechanism remains relatively simple. Future work may investigate adaptive and hierarchical fusion strategies to better balance semantic grounding, appearance preservation, and spatial consistency.

Second, UniCustom is mainly trained and evaluated at a resolution no larger than 512×512. While this setting is sufficient for validating the core design, it may limit the preservation of small text, logos, subtle facial details, and intricate textures. Scaling UniCustom to higher-resolution generation is

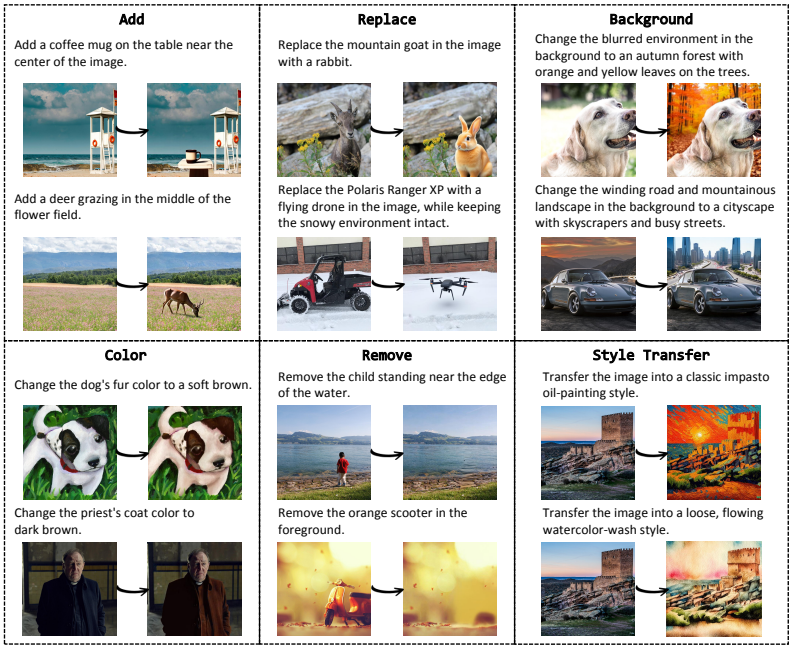

Figure 14: **Generated examples of UniCustom on image editing.**

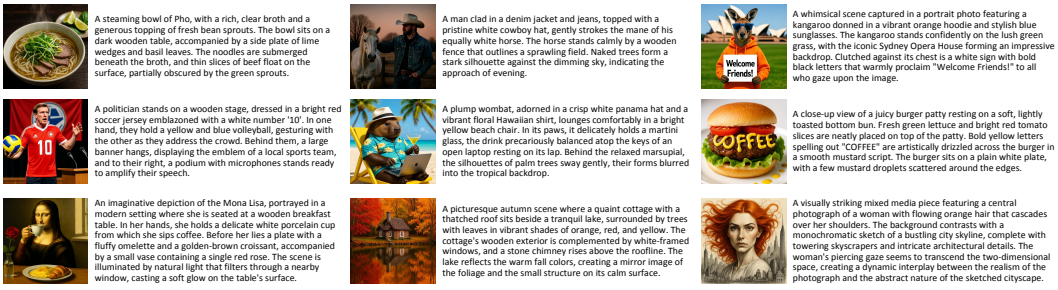

Figure 15: **Generated examples of UniCustom on text-to-image generation.**

therefore an important next step, especially for applications that require realistic identity preservation and fine-grained visual fidelity.

Third, UniCustom is primarily designed for multi-reference generation rather than as a fully general-purpose multimodal generation system. Although our two-stage training pipeline incorporates image understanding, image editing, and text-to-image data, these tasks mainly serve the central objective of reference-based generation. Specifically, image understanding data in pretraining helps retain the VLM's instruction-following and semantic grounding abilities under the unified visual representation, while editing and text-to-image data in supervised finetuning improve instruction adherence. Consequently, UniCustom exhibits certain generalization ability on auxiliary tasks, as shown by its multi-reference instruction understanding in Figure 6, image editing in Figure 14, and text-to-image generation in Figure 15. However, its performance on these tasks remains moderate and is not yet comparable to models specifically optimized for them.

These limitations arise from multiple aspects of the current training recipe. For understanding-oriented tasks, the pretraining stage includes only a small proportion of image understanding data, mainly captioning and general VQA. Although such data helps preserve the instruction-following and multimodal understanding abilities, it is insufficient for more complex multimodal reasoning tasks, such as mathematics and science. For generation-oriented auxiliary tasks, the editing and text-to-image data used during supervised finetuning are limited in both scale and diversity compared with those used by task-specialized models. As a result, UniCustom can generalize to image editing and text-to-image generation to some extent, but these capabilities remain secondary to its main reference-generation objective. In addition, the VLM is kept frozen throughout training, leaving the lightweight

fusion layer as the primary trainable component for adapting the unified visual representation. This design improves training efficiency and helps preserve the base VLM's capabilities, but it also constrains the model's adaptability to broader understanding and generation tasks.

Future work will explore higher-resolution generation, more expressive visual fusion modules, parameter-efficient or selective VLM adaptation, and improved training mixtures with more diverse and higher-quality data for understanding, editing, text-to-image generation, and reference-based generation. Overall, UniCustom represents an initial yet meaningful step toward unified visual conditioning for multimodal generation. Beyond its strong open-source multi-reference generation performance, our results suggest that fusing ViT- and VAE-derived visual representations before VLM encoding is a promising way to bridge understanding-oriented and generation-oriented visual signals. We hope this insight can support the development of a more general multimodal understanding and generation system.

G Ethical Statement

Our work focuses on algorithmic improvements for multi-reference image generation, aiming to improve subject consistency and compositional fidelity when synthesizing images from complex textual instructions and multiple reference images. The training data are mainly constructed from open-source datasets, with evaluation conducted on public benchmarks such as OmniContext [41] and MICo-Bench [39]. No data are collected from human subjects. Although we do not anticipate direct or immediate negative societal impacts from the algorithmic contributions themselves, we recognize that improvements in image generation quality and controllability can amplify existing risks associated with generative models. We therefore encourage responsible use of the proposed method and caution against deployment in contexts involving real identities, sensitive attributes, or high-stakes decision-making without additional safeguards.

H Reproducibility Statement

To ensure reproducibility, we provide detailed descriptions of the datasets, training setup, and hyperparameters in the main text and appendix. We will also release our code and checkpoints to further facilitate reproducibility soon.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main claims and contributions. These claims are consistent with the theoretical and experimental results, and are appropriately scoped to avoid overgeneralization.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Section 3.1 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code and model checkpoints are not yet publicly available. We plan to release them in the future, as stated in Appendix H.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: In Section 3.1 and Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We follow the widely-used evaluation benchmark, and these metrics do not require reporting error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted on a cluster of 128 GPUs as shown in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our work studies multi-reference image generation, which may pose potential misuse risks, such as generating misleading, impersonating, or otherwise harmful visual content from reference images. We therefore do not release any scraped or private data, and all training and evaluation data are constructed from open-source or public benchmark datasets subject to their respective licenses and usage terms. When the checkpoints are released, they will be distributed with usage restrictions, documentation of intended use, and safeguards against harmful applications, including prohibiting identity impersonation, non-consensual image generation, and generation of unsafe or deceptive content.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have carefully cited and stated the assets used in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets at submission time since the code, checkpoints, and processed training data are currently under institutional review. Appendix A and B documents the experimental setup and data sources, and we will release the assets together with usage instructions and documentation once the review process is completed.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The human evaluation details are described in Appendix C. Participants were recruited through CloudResearch and received monetary compensation for completing the evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: We have obtained IRB approvals from our institution.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing, editing, and formatting assistance, and did not affect the core methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.