Steering Visual Generation in Unified Multimodal Models with Understanding Supervision

Zeyu Liu[1,2,3,4*] **Zanlin Ni**[1,2,3] **Yang Yue**[1,2,3] **Cheng Da**[4] **Huan Yang**[4]
Di Zhang[4] **Kun Gai**[4] **Gao Huang**[1,2,3†]
[1]Department of Automation, Tsinghua University
[2]Beijing Key Laboratory of Embodied Intelligence Systems
[3]Institute for Embodied Intelligence and Robotics, Tsinghua University
[4]KlingAI Research
Project Page: `https://lzy-tony.github.io/uno`

Abstract

Unified multimodal models are envisioned to bridge the gap between understanding and generation. Yet, to achieve competitive performance, state-of-the-art models adopt largely decoupled understanding and generation components. This design, while effective for individual tasks, weakens the connection required for mutual enhancement, leaving the potential synergy empirically uncertain. We propose to explicitly restore this synergy by introducing ***Un**derstanding-**O**riented Post-Training* (**UNO**), a lightweight framework that treats understanding not only as a distinct task, but also a direct supervisory signal to steer generative representations. By incorporating objectives that encode semantic abstraction (captioning) and structural details (visual regression), we enable effective gradient flow from understanding to generation. Extensive experiments on image generation and editing demonstrate that understanding can serve as an effective catalyst for generation.

1 Introduction

Unified Multimodal Models (UMMs), which integrate language comprehension, visual understanding, and visual generation within a single framework, have recently achieved significant success [46, 10, 54, 48, 36, 9, 5]. By jointly modeling understanding and generation, these models facilitate versatile any-to-any interaction, enabling advanced and promising new capabilities that range from complex multimodal reasoning [26, 7] and free-form image manipulation [9] to interleaved world modeling [15, 58].

A long-term objective for Unified Multimodal Models is to achieve capability synergy [10, 48, 59], wherein multimodal understanding and generation do not merely coexist under a single framework, but mutually enhance one another. However, to maintain strong task-specific performance, state-of-the-art architectures increasingly adopt a decoupled representation paradigm [29, 56, 6, 34, 9], which aims to alleviate optimization conflicts between the high-level semantic abstraction required for understanding and the low-level objectives inherent to generative modeling. Concretely, these approaches separate understanding and generation into distinct representation spaces, ranging from distinct vision encoders [56] and feed-forward networks (FFNs) [27] to disjoint transformer parameters [29, 9]. While such decoupling effectively mitigates interference and preserves specialization, it inherently limits the extent to which the rich semantics learned by the understanding expert can be directly utilized by the generative components, leaving it less clear to what extent genuine capability synergy can be achieved within these frameworks.

*Work done while interning at KlingAI Research.
†Corresponding author.

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

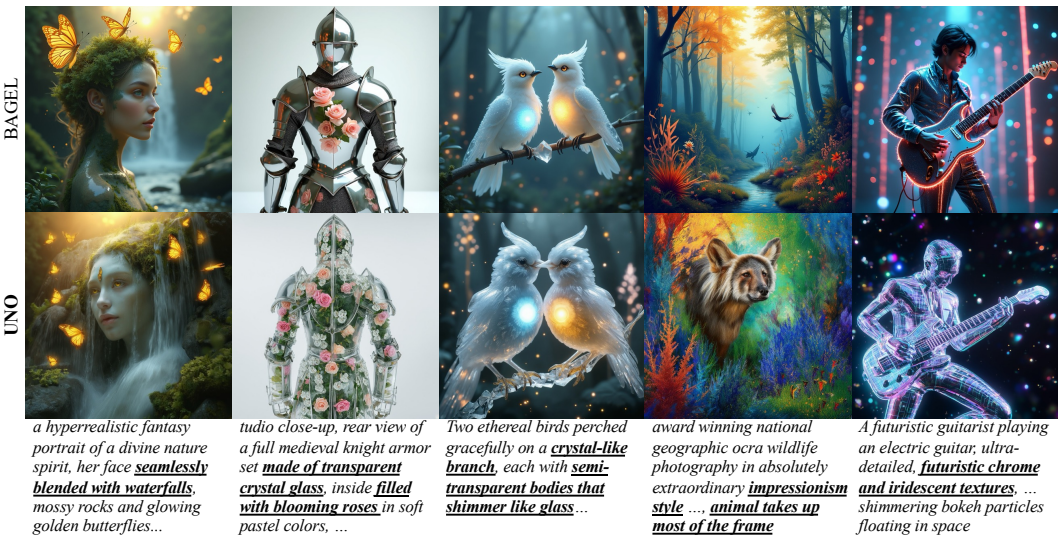

Figure 1: **Qualitative comparisons on image generation** between BAGEL and BAGEL-UNO.

In this work, we isolate this specific direction of synergy and investigate whether direct understanding-oriented supervision can be systematically leveraged to enhance generative learning in unified models. To this end, we propose ***Un**derstanding-**O**riented Post-Training* (**UNO**), a light-weight framework that explicitly supervises generative representations with understanding signals. Rather than treating understanding as a parallel task, we re-route the information flow by conditioning the frozen understanding expert on intermediate noised generative representations, strengthening direct gradient flow from understanding to generation. Specifically, we incorporate two complementary understanding-oriented proxy objectives for optimizing generative representations: (i) language supervision via captioning and (ii) visual understanding supervision via regressing with metaquery tokens. Language supervision enhances discriminative concepts through high-level abstraction, yet is inherently sparse and may overlook fine-grained details. In contrast, visual understanding supervision captures denser details and spatial structures, providing the structural information that abstract linguistic signals lack. By integrating these complementary objectives, UNO enriches the model's generative representations with multimodal semantics without increased architectural complexity.

Building on this conceptual framework, we conduct a systematic evaluation across diverse generation tasks. Extensive experiments across image generation and editing benchmarks with two different backbones indicate that UNO yields consistent and substantial improvements over strong baselines without degrading understanding performance. Specifically, UNO improves the competitive BAGEL [9] baseline on both image generation (GenEval2 71.7 $\rightarrow$ 75.1, DPG-Bench 84.03 $\rightarrow$ 86.12, UniGenBench++ 61.53 $\rightarrow$ 65.03) and image editing tasks (GEdit-Bench-EN 6.52 $\rightarrow$ 7.17, GEdit-Bench-CN 6.50 $\rightarrow$ 7.20) by significant margins. Beyond quantitative gains, qualitative visualizations further reveal improved semantic structure for generative representations at heavily noised timesteps. These results demonstrate that in unified models, strong multimodal understanding can be harnessed to directly benefit generation, paving the way for more integrated unified multimodal systems.

2 Related Work

2.1 Unified Multimodal Models

Inspired by the success of large language models (LLMs) [1, 50, 16, 67] and advances in separate multimodal understanding [32, 53, 2] and generation [39, 38, 37, 11, 3] systems, recent works have moved toward unified multimodal models that perform both multimodal understanding and generation within a unified framework. Early approaches often relied on quantized autoregressive modeling of visual content [46, 54, 56, 6, 61], *i.e.*, transforming images into a sequence of tokens with discrete vector quantizers and processing those tokens autoregressively in a way akin to language modeling. While these methods demonstrated the feasibility of unified modeling, their generative quality is constrained by the discretization bottleneck imposed by VQ-based tokenizers. To overcome these limitations, recent approaches combine multimodal large language models (MLLMs) for

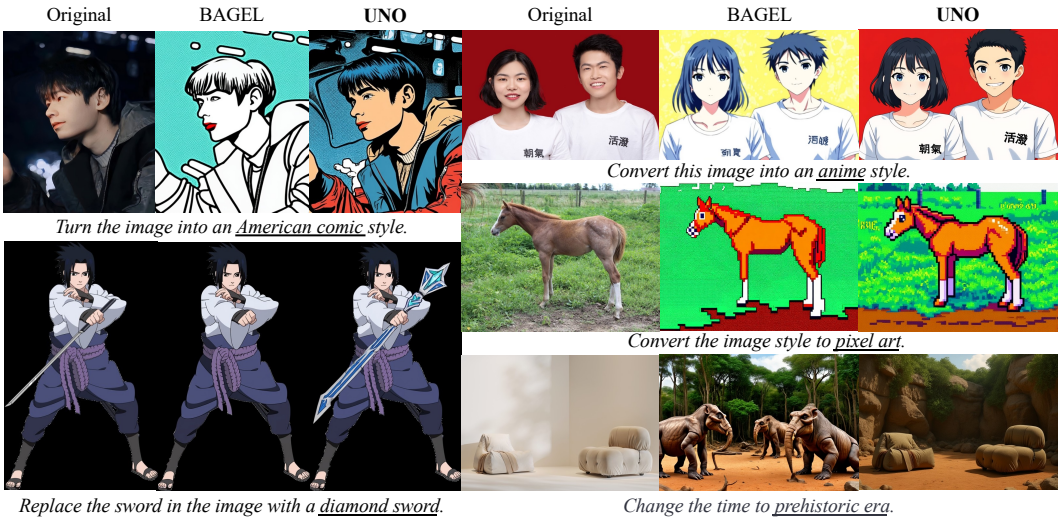

Figure 2: **Qualitative comparisons on image editing** between BAGEL and BAGEL-UNO.

understanding with diffusion models for generation, yielding substantially improved expressivity and performance. Within this hybrid paradigm, one thread of research arranges an MLLM backbone sequentially with a diffusion decoder. Implementations include either predicting through special query tokens [44, 10, 13, 36, 60], or through predicting intermediate latent representations [48, 5] that are consumed by the diffusion-based generator. A complementary line of work emphasizes parallel architectures that process understanding and generation within a unified backbone. These designs include integrated transformer [34, 66] as well as Mixture-of-Experts [27] or Mixture-of-Transformers [29, 30, 9] formulations that allocate capacity across modalities and tasks.

2.2 Representations in Unified Multimodal Models

A central challenge in unified multimodal modeling lies in representation design. Unified models must simultaneously support multiple, potentially conflicting tasks, each imposing distinct requirements on the underlying representations. While early unified models enforced a single shared representation for all visual signals [46, 54], subsequent studies have shown that such designs lead to potential task conflicts [65] that degrade task-specific performance. As a result, contemporary models increasingly adopt *decoupled visual representations* to better accommodate divergent objectives. One common strategy is to employ separate vision encoders for understanding and generation [56, 6, 66]. Beyond encoder decoupling, more recent architectures further separate representations within the backbone itself. Mixture-based designs, including MoE and MoT [29, 30, 9, 27], explicitly allocate distinct pathways to understanding and generation. This suggests that unified models operate over multiple representations and that effective coordination among these representations is critical for performance.

2.3 Understanding Priors for Generation

Recent studies have increasingly highlighted the importance of understanding-oriented priors in improving generative modeling. Representation alignment methods such as REPA [69] and REPA-E [25] regularize diffusion training by aligning intermediate features with pretrained semantic visual representations, substantially accelerating convergence and performance. Beyond alignment losses, RAE [71, 49] and SVG [42, 41] redesign latent spaces around semantically rich encoder representations, enabling high-quality generation without relying on traditional VAEs.

3 Approach

3.1 Preliminary: Information Flow and Representations in Unified Multimodal Models

Representative state-of-the-art unified multimodal models, *e.g.* BAGEL [9], are typically initialized from pretrained vision–language models (VLMs) and comprise specialized experts for understanding and generation. Visual understanding is handled exclusively by the understanding expert, which

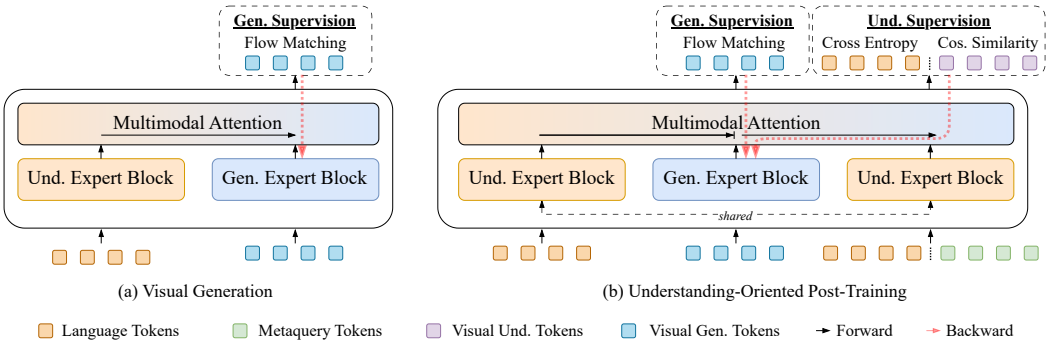

Figure 3: **Conceptual illustration of training process and backward gradient flow**. (a) *Generation Training*: Current generative training in unified models encode conditions using the understanding expert and transfers information uni-directionally via conditioning to the generative expert, where the outputs are optimized using low-level flow matching objectives. Generation experts receive gradients solely from generative targets, while signals from the understanding pathway remain isolated. (b) *Understanding-Oriented Post-Training*: Understanding-oriented post-training jointly supervises the sample with generation and understanding objectives. Specifically, the understanding expert conditions on the noised generative representations, and jointly conducts (i) language supervision through captioning and (ii) visual understanding supervision by using metaquery tokens to predict subsequent understanding tokens. This enables generation blocks to receive gradients from both pathways, enabling strong understanding supervision to directly shape generative representations.

jointly processes visual understanding and language tokens in isolation from the generation pathway. Conversely, visual generation is conditioned on representations encoded by the understanding expert and supervised by low-level flow-matching objectives, as illustrated in Figure 3(a). As a result, although unified models consolidate diverse capabilities within a single architecture, their internal representations adopt decoupled designs and exhibit distinct characteristics, as summarized in Table 1. Language representations are highly abstract but lack dense information arising from visual details; conversely, generation representations are visually dense but often lack rigorous semantic organization. Visual understanding representations occupy an intermediary position, retaining visual granularity while encoding structured semantics.

3.2 Motivation

Table 1: **Different representations** in UMMs (*e.g.* BAGEL [9]).

Representation	Abstraction	Density
Language	High	Low
Visual Und.	Medium	Medium
Visual Gen.	Low	High

As previously established, the decoupled architecture induces a unidirectional information flow from understanding to generation. Although the generation expert is conditioned on the understanding expert and can therefore inherit semantic cues implicitly, the generative flow-matching objective provides only weak supervision for enforcing semantic structure in the generative representations. Consequently, a pronounced performance gap emerges: while the understanding expert exhibits strong semantic capabilities, the generation expert frequently struggles with complex instructions and fine-grained semantic adherence. This disparity indicates that, in current unified frameworks, understanding capabilities are substantially stronger than generation, yet remain largely under-exploited as a source of supervision. Importantly, unified models already encode rich semantic representations through language and visual understanding, therefore we argue that relying solely on low-level objectives is suboptimal for training the generative pathway. Motivated by this observation, we hypothesize that explicitly supervising generative representations through understanding objectives operated by the model's own understanding expert can inject strong semantic constraints, yielding a more semantically grounded representation space and ultimately improved generative performance.

3.3 Language and Visual Understanding Supervision

Language Supervision To address this, we first exploit the strong language-based understanding capabilities and re-route the information flow to enable *language supervision* directly over generated visual representations, as conceptually illustrated in Figure 3(b). Rather than conditioning the

understanding expert on visual understanding tokens, we condition it on the noised generation representations from the generation expert. The understanding expert then processes these features to output language tokens, supervised by an image captioning objective. This forces the understanding expert to decode semantics directly from the intermediate generation representations, effectively distilling abstract pretrained semantics into the generative pathway through backward gradient flows from understanding to generation. To preserve pretrained capabilities, we freeze the understanding expert, compelling the generation representations to adapt to the understanding-oriented objective.

A critical challenge in this setup is avoiding trivial solutions from information leakage. To mitigate this, we mask conditional prompt tokens when forwarding supervision language tokens, as shown in Figure 4. However, we empirically observe that supervising generated images with captions derived from the original prompt leads to abnormally low captioning loss. We hypothesize that the high information capacity of visual representations allows the model to "store" low-density prompt signals, creating shortcuts where the understanding expert simply copies rather than performing genuine semantic extraction.

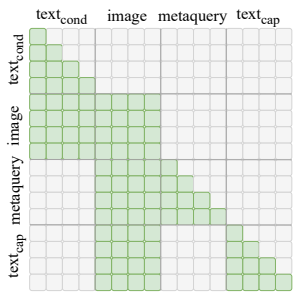

Figure 4: **Attention mask** for packed text-to-image training sample sequence.

To mitigate this, we adopt semantic augmentations by re-captioning the images using alternative captioning models. Target captions are semantically consistent but lexically different from the original conditioning prompts. By supervising the model with non-token-aligned captions, we force the understanding expert to rely on extracting semantic content from generation representations rather than surface-level token copying. The resulting objective is defined as:

$$\mathcal{L}_{\text{language}} = -\sum_{i=1}^{L} \log p(\mathbf{z}_i | \mathbf{z}_{<i}, \mathbf{V}_{\text{gen}}), \tag{1}$$

where $\mathbf{z}$ denotes the supervision tokens, and $\mathbf{V}_{\text{gen}}$ represents the noised visual representations.

Visual Understanding Supervision While language supervision provides high-level semantic guidance, language-based captions are inherently limited in information density: they often omit fine-grained visual details and cannot fully describe all aspects of an image. Moreover, language supervision lacks explicit 2D semantic structure encoded in visual based representations. To address these limitations, we introduce *visual understanding supervision* that complement language-based signals.

Specifically, we use the pretrained understanding expert as a strong visual prior. As the understanding expert is not designed to directly produce visual features, we adopt the MetaQuery framework [36] and insert a set of learnable metaqueries into the understanding expert. These metaqueries are processed autoregressively, and their output hidden states are supervised to regress dense visual features extracted from the target image by the model's native visual encoder [51]. This yields the visual supervision objective:

$$\mathcal{L}_{\text{vision}} = -\sum_{j=1}^{N} \text{sim}(\mathbf{v}_j, \mathbf{h}_j), \tag{2}$$

where $\mathbf{v}_j$ and $\mathbf{h}_j$ denotes the target representations from the pretrained visual encoder and the output states of the metaqueries, respectively. $\text{sim}(\cdot)$ represents cosine similarity.

Joint Supervision As summarized in Table 1, visual understanding representations contains fine-grained details and explicit 2D spatial structure that enrich language supervision. Conversely, language captions offer a more abstract and direct supervision, providing complementary high-level semantic guidance to visual supervision. Together, these two forms of supervision enable a more comprehensive learning signal. To enable joint supervision, we combine the proposed objectives with the standard flow-matching loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mse}} + \lambda_1 \mathcal{L}_{\text{language}} + \lambda_2 \mathcal{L}_{\text{vision}}, \tag{3}$$

To maximize training data efficiency, we employ a unified data packing strategy that concatenates all supervision signals into a single sequence. As shown in Figure 4, we modify the attention mask to manage information flow and avoid leakage between these tasks. This end-to-end configuration forces the generative pathway to optimize for both generation and understanding signals simultaneously.

Table 2: **Quantitative image generation performance**. †: evaluation with self-CoT. *: locally evaluated results. Know.: World knowledge. Attr.: Attribute. Act.: Action. Relat.: Relationship. Rea.: Logical reasoning. Gram.: Grammar. Comp.: Compound. Lay.: Layout. Ovr.: Overall.

	Model	GenEval2	DPG-Bench	UniGenBench++										
				Style	Know.	Attr.	Act.	Relat.	Rea.	Gram.	Comp.	Lay.	Text	**Ovr.**
Gen. Only	SDXL	45.3	74.65	87.40	72.63	44.34	34.22	44.92	9.55	47.33	26.68	29.85	1.15	39.75
	SD-3.5-Medium	64.8*	83.79*	89.80	84.34	66.99	60.65	68.78	37.73	59.89	53.35	70.34	15.23	60.71
	SD-3.5-Large	67.8	83.86*	88.60	88.92	68.59	62.17	69.80	32.27	58.96	58.76	69.03	32.76	62.99
	FLUX.1-dev	67.0	84.00	83.90	88.92	67.84	62.17	67.26	30.91	60.96	47.04	71.83	32.18	61.39
	Infinity	61.2*	83.46	90.80	87.97	68.06	60.17	69.16	31.36	60.16	51.42	66.60	12.36	59.81
	OmniGen2	67.4*	83.57	91.90	86.39	72.12	62.83	68.27	32.50	59.89	56.31	71.64	29.02	63.09
	Wan2.2-t2i-plus	-	-	91.10	87.34	70.19	68.00	73.03	42.05	66.53	61.37	74.77	13.83	64.82
Unified	Janus	53.7*	79.68	89.90	73.58	54.81	50.38	55.08	26.82	59.09	46.65	54.85	1.15	51.23
	Janus-Pro-7B	54.5*	84.19	90.80	86.71	67.74	64.26	68.40	37.05	64.44	62.11	72.01	2.59	61.61
	Emu3-8B	41.2*	80.60	86.80	77.06	51.39	40.11	49.75	19.32	52.94	36.86	44.78	1.15	46.02
	OneCAT-3B	60.7*	84.53	93.30	82.28	63.46	58.56	68.15	33.41	60.83	56.96	64.74	1.15	58.28
	Janus-Flow	42.7*	-	86.20	62.50	47.97	43.35	50.00	21.14	60.29	45.10	46.46	0.86	46.39
	BLIP3-o-8B	59.8*	81.60	92.80	80.22	63.89	63.97	66.50	39.55	68.58	53.74	68.47	1.15	59.87
	UniWorld-V1	70.0*	-	91.10	82.91	70.62	67.21	67.13	38.41	63.77	54.51	69.03	26.44	63.11
	Mogao-7B	-	84.33	-	-	-	-	-	-	-	-	-	-	-
	BAGEL	71.7†	84.03	90.20	85.60	67.74	61.98	70.69	30.23	66.44	58.12	76.49	7.76	61.53
	BAGEL + UNO	**75.1**†	**86.12**	92.30	84.18	72.86	67.49	75.25	33.26	68.98	64.82	80.78	10.34	**65.03**
	Δ	**+3.4**	**+2.09**	**+2.10**	**-1.42**	**+5.12**	**+5.51**	**+4.56**	**+3.03**	**+2.54**	**+6.70**	**+4.29**	**+2.58**	**+3.50**

4 Experiments

4.1 Experiment Setup

Training For all experiments, we train BAGEL-7B [9] for 5K iterations while keeping the understanding expert frozen. For the image generation task, we utilize a curated set of high-quality text-image pairs. Notably, we exclude distillation-based data such as BLIP3o-60k [5] to prevent evaluation template leakage for certain evaluation benchmarks [64]. We apply semantic augmentation and re-caption images using different caption models to form text-image-text triplets. For image editing, training is conducted on CrispEdit-2M [8], a diverse set of high quality editing pairs. We additionally caption the target images to provide supervision text, resulting in text-image-image-text quadruplets.

Evaluation Protocol We evaluate text-to-image generation performance using GenEval2 [20], DPG-Bench [18] and UniGenBench++ [55]. We primarily focus on DPG-Bench for its diverse prompts to evaluate semantic related instruction following. Additionally, as DPG-Bench exhibit rapid performance saturation [45], we also evaluate on UniGenBench++, a more recent and fine-grained evaluation set. We also report results on GenEval2 [20], an improved version of GenEval [14] that mitigates evaluation errors and drifts from human judgment. Evaluations on DPG-Bench and UniGenBench++ are conducted without activating thinking mode or prompt rewriting, while GenEval2 is evaluated with CoT following default setting in [20]. We evaluate with 4 random seeds to balance robustness and computation costs. For image editing, we evaluate on GEdit-Bench-EN/CN [33], a comprehensive multilingual benchmark derived from real-world user instructions.

Baselines We compare against both generation-only and unified models. For image generation, generation-only baselines include SDXL [38], Stable Diffusion 3.5 Medium/Large [11], FLUX.1-dev [3], Infinity [17], OmniGen2 [57] and Wan2.2-t2i-plus, unified models include Janus [56], Janus-Pro [6], Emu3 [54], OneCAT [27], Janus-Flow [34], BLIP3-o [5], UniWorld-V1 [31], Mogao [30] and BAGEL [9]. For editing, generation-only models include Instruct-Pix2Pix [4], MagicBrush [70], AnyEdit [19], OmniGen [63], OmniGen2 [57], Step1X-Edit [33] and FLUX-Kontext [24], and unified models include BAGEL [9], BAGEL-NHR [22] and UniWorld-V1 [31].

4.2 Main Results

Image Generation We report generation performance in Table 2. UNO yields consistent improvements over the original BAGEL as well as generation-only and unified-model baselines across GenEval2, DPG-Bench, and UniGenBench++. On UniGenBench++, UNO shows pronounced gains



Figure 5: Qualitative visualizations of image generation results.

in dimensions including compound, attribute, action, and relationship, which are closely related to semantic comprehension. We observe only a slight degradation in world-knowledge scores, indicating that it is more dependent on the diversity and coverage of the training data.

Image Editing We report quantitative image editing results on GEdit-Bench-EN/CN in Table 3. UNO consistently improves editing performance over BAGEL and other strong baselines across semantic consistency, perceptual quality, and overall metrics. On GEdit-Bench-EN, UNO achieves the best overall score, with notable gains in perceptual quality while preserving edit intent. Improvements also transfer to Chinese evaluations on GEdit-Bench-CN despite training solely on English data, demonstrating robust generalization.

Table 3: **Quantitative image editing performance**. G_SC, G_PQ and G_O denotes GPT-4.1 evaluated semantic consistency, perceptual quality and overall performance. We mark the best results for each dimension in **bold** and underline the second best.

Model	GEdit-Bench-EN			GEdit-Bench-CN		
	G_SC	G_PQ	**G_O**	G_SC	G_PQ	**G_O**
Instruct-Pix2Pix	3.58	5.49	3.68	-	-	-
MagicBrush	4.68	5.66	4.52	-	-	-
AnyEdit	3.18	5.82	3.21	-	-	-
OmniGen	5.96	5.89	5.06	-	-	-
OmniGen2	7.16	6.77	6.41	-	-	-
Step1X-Edit	7.09	6.76	6.70	7.20	6.87	6.86
FLUX-Kontext	6.95	7.30	6.27	-	-	-
UniWorld-V1	4.93	7.43	4.85	-	-	-
BAGEL-NHR	**8.04**	6.87	7.08	-	-	-
BAGEL	7.36	6.83	6.52	7.34	6.85	6.50
BAGEL + UNO	7.76	**7.54**	**7.17**	**7.80**	**7.54**	**7.20**
Δ	**+0.40**	**+0.71**	**+0.65**	**+0.46**	**+0.69**	**+0.70**

Qualitative Results. We present qualitative comparisons of image generation and editing between the original BAGEL and UNO in Figure 1 and Figure 2, respectively. A more comprehensive comparison is presented in Section E of the appendix. For generation, UNO demonstrates stronger instruction following. For editing, it more effectively interprets abstract instructions and better preserves fine-grained background details, benefiting from the additional understanding supervision during training. Further qualitative examples are provided in Figure 5 and Figure 6. Full prompt list is displayed in Table 19 of the appendix.

4.3 Analysis

UNO as an effective post-training approach To assess the effectiveness of UNO as a post-training strategy, we decompose UNO and compare it against other tuning-based post-training approaches under identical data settings. For image generation, we compare with supervised fine-tuning (SFT) and reconstruction alignment (RecA [64]), and present results in Table 4. Our observations are threefold: 1) applying either language or visual understanding supervision alone consistently improves performance over SFT, with language supervision exhibiting most pronounced gains. 2) Joint supervision yields further gains beyond single-modality supervision, indicating a complementary relationship language and visual supervision. 3) Joint supervision outperforms both SFT and RecA, demonstrating its effectiveness.

For image editing, we conduct analogous ablation studies and decompose the understanding objective into its individual components, where the setting without language or visual supervision corresponds

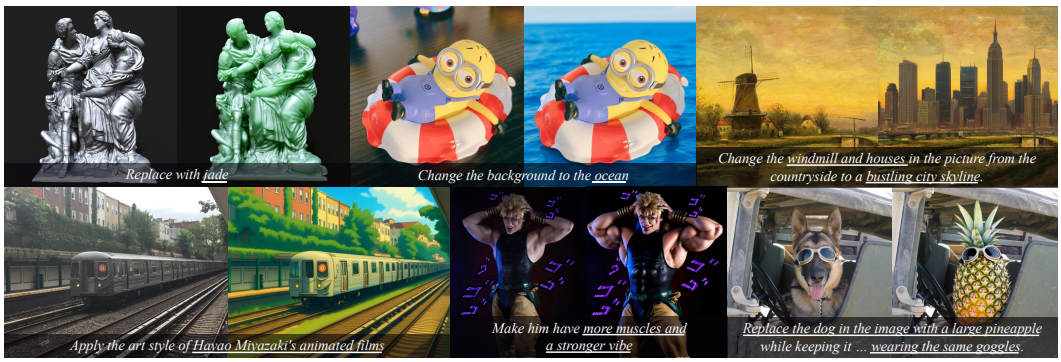

Figure 6: Qualitative visualizations of image editing results.

to SFT on the same data. As reported in Table 5, each supervision signal independently improves upon the SFT baseline, while their combination achieves the best performance on GEdit-Bench. These results mirror the image generation findings and further confirm the effectiveness and complementary benefits of language and visual understanding supervision.

UNO improves generative features Following prior feature-space analyses [21, 25], we investigate the effect of UNO on generative representations by visualizing noised features from the generation expert. Specifically, we apply PCA to intermediate representations at early stages of denoising, as shown in Figure 7. When trained with generative supervision alone, representations exhibit substantial noise and weak semantic organization (*e.g.* ambiguity in Col. 1), failing to preserve fine-grained visual details (*e.g.*, missing gift box at the bottom-left in Col. 2, wrong pose in Col. 3). In contrast, UNO reduces noise and improves the discriminability of visual details, suggesting that understanding supervision encourages the model to form more semantically grounded and robust generative representations.

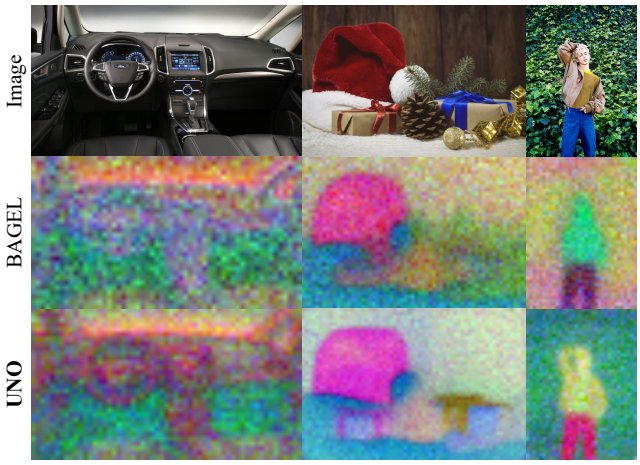

Figure 7: **Visualization of latent features**. We visualize the structure of latent features of the generation expert at highly noised timesteps. Empirically, we observe that understanding supervision improves latent structures, reduces noise while preserving better semantic information and details.

Optimization gradient directions To investigate the effect of understanding supervision on optimization and gradient dynamics in generative training, we visualize the per-layer gradient directions induced by the understanding and generative objectives in Figure 8. We observe that gradients are largely orthogonal across most layers, with a subset exhibiting positive alignment. This suggests that, the understanding objective does not introduce optimization conflicts with denoising and instead provides complementary semantic guidance under UNO.

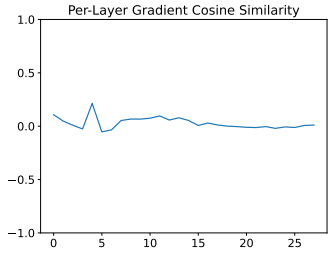

Figure 8: **Visualization of gradient similarity** between understanding and generation objectives.

Semantic augmentation for language supervision We ablate the effect of semantic augmentation with different captions in Table 7. Empirically, we discover that augmentation effectively reduces leakage and enhances performance. We also discover that even though using the same prompt is prone to leakage, it can still improve generation.

Vision connector design and more We further study how different vision connector and supervision targets affect performance. We adopt a wide range of established approaches, including ViT-based

Table 4: **Comparison between different supervision methods for post-training BAGEL**. For a fair comparison, we also implement RecA [64] on our training dataset, denoted by †.

Model	GenEval2	DPG-Bench	UniGenBench++
BAGEL	71.7	84.03	61.53
w/ RecA	74.9	85.29	63.74
w/ RecA†	74.3	85.37	64.12
w/ SFT	73.6	84.91	63.02
w/ Language	74.6	85.69	64.38
w/ Visual Und.	74.0	85.31	63.60
w/ Joint	**75.1**	**86.12**	**65.03**

Table 5: **Effect of different supervision targets**. We decompose the joint supervision objective and ablate the individual contributions on image editing. Double cross mark denotes the SFT baseline trained without UNO. Lang. and Vis. denotes language and visual understanding supervision respectively.

Lang.	Vis.	GEdit-Bench-EN			GEdit-Bench-CN		
		G_SC	G_PQ	**G_O**	G_SC	G_PQ	**G_O**
✗	✗	7.61	7.44	7.00	7.63	7.42	6.96
✗	✓	7.69	7.47	7.09	7.77	7.54	7.18
✓	✗	7.66	7.47	7.09	7.69	7.54	7.11
✓	✓	**7.76**	**7.54**	**7.17**	**7.80**	**7.54**	**7.20**

Table 6: **Effect of connector and supervision targets** for visual supervision. Sim.: cosine similarity. De.: Denoise. Default marked in gray.

Arch.	Target	GenEval2	DPG-Bench	UniGenBench++
Identity	Sim.	**74.0**	**85.31**	63.60
ViT	Sim.	73.7	85.21	63.47
MLP	De.	73.7	85.09	63.41
DiT	De.	73.8	85.09	**63.82**

Table 7: **Effect of language augmentation**. Default in gray.

Aug.	GenEval2	DPG-Bench	UniGenBench++	Loss
✗	74.1	85.26	64.09	∼0.1
✓	**74.6**	**85.69**	**64.38**	∼0.7

similarity [36], adaLN-MLP–style denoising [28, 47], and DiT-based denoising reconstruction [52]. As summarized in Table 6, we empirically observe that different architectures and targets results in similar performance for generation. Therefore, we adopt the most simple design using identity projection and cosine similarity. Further ablations are presented in Section D in the appendix.

General applicability To further assess the generality of UNO, we conduct experiments on Show-o [65], an alternative unified model that integrates VLM with discrete diffusion. We train Show-o using the processed generation data with understanding data from LLaVA [32] to preserve understanding capabilities. We report results across generation and understanding (MME [12]) benchmarks in Table 8. As demonstrated, UNO successfully improves upon the Show-o baseline across generation benchmarks while preserving understanding performance, demonstrating UNO's generality.

Table 8: Effect of UNO on Show-o.

Model	GenEval2	DPG-Bench	UniGenBench++	MME
Show-o	58.57	80.04	50.39	1233
+UNO	**61.26**	**81.94**	**53.41**	**1238**

UNO and CoT Self-CoT is a commonly adopted strategy for improving generation performance in unified models. We conduct a systematic analysis of the interplay between CoT and UNO. Our results in Table 9 show that UNO consistently outperforms CoT alone, while also serving as a complementary mechanism that further enhances CoT-based generation when combined.

Table 9: Effect of UNO and CoT on BAGEL.

CoT	UNO	GenEval2	DPG-Bench	UniGenBench++
✗	✗	69.6	84.03	61.53
✗	✓	72.0	**86.12**	**65.03**
✓	✗	71.7	83.69	61.36
✓	✓	**75.1**	85.33	64.95

Limitations and future work UNO is a general-purpose framework that does not leverage specialized data for vertical domains. Therefore, on tasks requiring such capabilities (*e.g.*, knowledge retrieval in WISE [35]), it remains largely orthogonal. Preliminary results indicate these capabilities are complementary and do not conflict with our design. To verify this, we construct 50k distilled knowledge-oriented samples and continue training UNO for 1k steps, yielding consistent gains (Table 10). We will further explore improving on these specialized domains in future work.

Table 10: **Evaluation on WISE**. * indicates additional training with augmented knowledge data.

Model	Cult.	Time	Spa.	Bio.	Phy.	Chem.	Ovr.
BAGEL	0.44	0.55	0.68	0.44	0.60	0.39	0.52
+UNO	0.44	0.52	0.63	0.45	0.57	0.42	0.51
+UNO*	0.49	0.56	0.64	0.52	0.67	0.46	**0.56**

5 Conclusion

In this work, we present UNO, a preliminary exploration into supervising generation directly with understanding in unified multimodal models. Through designing a light-weight post-training framework that strengthens gradient flow from understanding to generation via a combination of complementary objectives, we demonstrate that understanding can serve as effective catalyst to enhance generation. Extensive experiments across different generative tasks validate the effectiveness of UNO. Our findings show promising possibilities for more integrated unified models with deeper synergy.

Acknowledgements

This work is supported in part by the National Key R&D Program of China under Grant 2024YFB4708200, the National Natural Science Foundation of China under Grants U24B20173, U2541227 and 62625604, the Lenovo Academic Collaboration Project, and the Scientific and Technological innovation action plan of Shanghai Science and Technology Committee (No.22511101502). We would also like to extend our deepest appreciation to Jiayi Guo, Yifan Pu and Xu Zhang for insightful discussions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[3] BlackForest. Black forest labs; frontier ai lab, 2024.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[5] 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.

[6] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[7] Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025.

[8] Wei Chow, Linfeng Li, Lingdong Kong, Zefeng Li, Qi Xu, Hang Song, Tian Ye, Xian Wang, Jinbin Bai, Shilin Xu, et al. Editmgt: Unleashing potentials of masked generative transformers in image editing. *arXiv preprint arXiv:2512.11715*, 2025.

[9] 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.

[10] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[13] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[14] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[15] Chenhui Gou, Zilong Chen, Zeyu Wang, Feng Li, Deyao Zhu, Zicheng Duan, Kunchang Li, Chaorui Deng, Hongyi Yuan, Haoqi Fan, et al. Vq-va world: Towards high-quality visual question-visual answering. *arXiv preprint arXiv:2511.20573*, 2025.

[16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[17] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025.

[18] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

[19] Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Mingyang Wan, Guojun Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Anyedit: Edit any knowledge encoded in language models. In *Forty-second International Conference on Machine Learning*.

[20] Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

[21] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*.

[22] Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. Nohumansrequired: Autonomous high-quality image editing triplet mining. *arXiv preprint arXiv:2507.14119*, 2025.

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[24] 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.

[25] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

[26] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Forty-second International Conference on Machine Learning*.

[27] Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025.

[28] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024.

[29] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[30] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

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

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[33] 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.

[34] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025.

[35] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[36] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

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

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[41] Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, et al. Svg-t2i: Scaling up text-to-image latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2512.11749*, 2025.

[42] Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301*, 2025.

[43] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678, 2023.

[44] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[45] Bingda Tang, Boyang Zheng, Sayak Paul, and Saining Xie. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28586–28595, 2025.

[46] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[47] NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, et al. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025.

[48] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

[49] Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

[50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[51] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[52] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[54] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[55] Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025.

[56] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025.

[57] 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.

[58] Jialong Wu, Xiaoying Zhang, Hongyi Yuan, Xiangcheng Zhang, Tianhao Huang, Changjing He, Chaoyi Deng, Renrui Zhang, Youbin Wu, and Mingsheng Long. Visual generation unlocks human-like reasoning through multimodal world models. *arXiv preprint arXiv:2601.19834*, 2026.

[59] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *International Journal of Computer Vision*, 134(1):39, 2026.

[60] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025.

[61] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[62] Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang YU, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[63] 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.

[64] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025.

[65] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*.

[66] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[67] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[68] 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[69] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*.

[70] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[71] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A Training Settings

We present the detailed training parameters for training image generation and editing tasks in Table 11. All experiments are trained with compute equivalent to approximately 32 A100 GPU days.

Table 11: Detailed hyper-parameters for post-training BAGEL on image generation and editing tasks.

Hyper-parameter	Image Generation	Image Editing
Iterations	5K	5K
#Samples	$\sim$2M	$\sim$750K
Learning Rate	0.0001	0.0001
Schedule	Cosine	Constant
Weight Decay	0.0	0.0
λ_1	0.1	0.1
λ_2	0.2	0.2
EMA ratio	0.9999	0.9999
#Metaqueries	256	256
Grad Norm Clip	1.0	1.0
Timestep Shift	4.0	4.0
Resolution	(512, 1024)	(512, 1024)

B Training Sample Statistics

We report the average token count for each component in the packed training samples of UNO. As shown in Table 12, UNO introduces only a marginal token overhead (text captions and meta-query tokens) compared to standard text-to-image training (text conditions and image tokens).

Table 12: Average token count for each component when training UNO.

Text Cond.	Img	Text Cap.	Metaquery
171	3362	80	256

C Image Generation Training Data

We curate our dataset from a diverse set of open-source resources, including LAION [40], JourneyDB [43], and OpenImages [23], etc. The collected data spans a wide range of categories, encompassing natural, synthetic, and design imagery, with rich subcategories such as objects, landscapes, human subjects, animals, plants, food, indoor scenes, artistic designs, and sports. Following data collection, we apply a systematic filtering pipeline based on resolution (excluding images below 512 pixels), aesthetic quality, visual clarity, color saturation, and safety considerations. Finally, we employ an internal captioning model to produce detailed, dense textual descriptions, yielding prompts with an average length of approximately 200 tokens.

For the recaptioning pipeline, we adopt a simple and straightforward approach, prompting Qwen2.5-VL-7B with the target image, and the system prompt "You are an ai captioning assistant, please describe this image in detail. Be absolutely accurate with your caption, do not imagine, hallucinate or hint at contents that is NOT present in the image."

D Further Ablations

Number of Metaqueries We study the impact of varying the number of metaqueries used for visual supervision, with results reported in Table 13. As shown, increasing the number of metaqueries beyond 256 leads to a degradation in performance. This trend suggests that visual understanding supervision at a resolution of 224×224 is sufficient to capture the necessary semantic details for effectively supervising the generation process.

Unfreezing the Understanding Expert We ablate the effect of jointly training the understanding expert, with results summarized in Table 14. We find that unfreezing and fine-tuning the understanding expert does not yield consistent improvements in generation performance. Moreover, optimizing

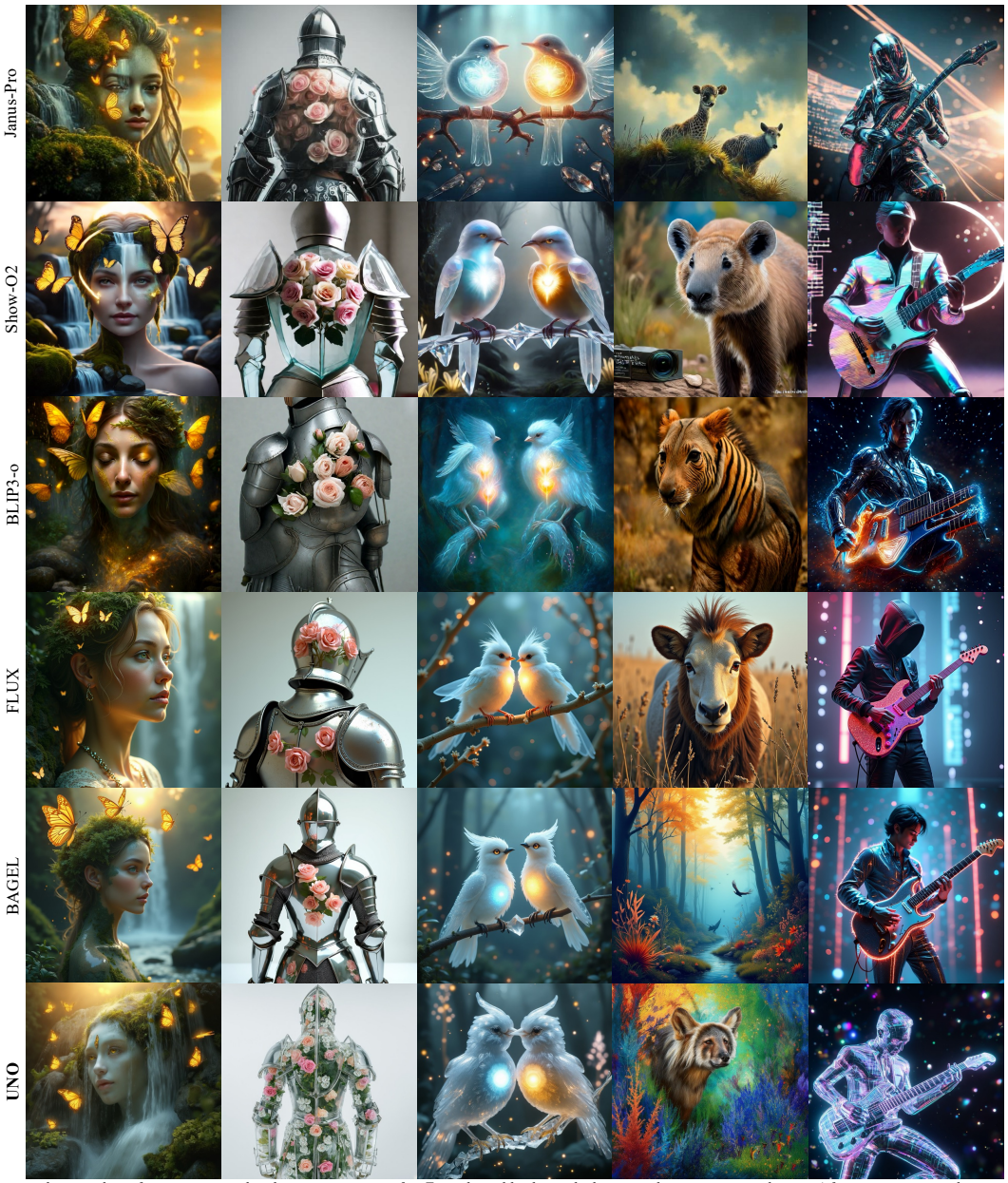

Figure 9: **More complete qualitative comparisons on image generation** between UNO and competitive generation and unified model baselines.

the understanding expert solely with the proxy objective may degrade its performance on standard understanding benchmarks, which can in turn weaken the quality of the supervision signals provided to the generation expert.

Effect of masking condition prompts We investigate the effect of masking the condition prompts and present results in Table 15. We observe that without masking, information leakage mostly undermines the effect of understanding supervision.

Effect of causal prediction for visual supervision We investigate the effect of masking the condition prompts and present results in Table 16. We observe that predicting the metaquery tokens in causal or bidirectional order does not significantly affect performance.

Table 13: **Effect of different number of metaqueries**. Default marked in gray.

#Metaqueries	GenEval2	DPG-Bench	UniGenBench++
256	74.0	85.31	63.60
576	73.4	85.08	63.36

Table 14: **Effect of unfreezing the understanding expert**. Default marked in gray.

Train Und.	GenEval2	DPG-Bench	UniGenBench++
✗	75.1	86.12	65.03
✓	75.8	84.62	64.36

Table 15: **Effect of masking conditional prompts**. Default marked in gray.

Mask	GenEval2	DPG-Bench	UniGenBench++
✗	73.4	84.97	62.76
✓	75.1	86.12	65.03

Table 16: **Effect of prediction order for visual supervision**. Default marked in gray.

	GenEval2	DPG-Bench	UniGenBench++
Causal	75.1	86.12	65.03
Bidirectional	75.3	85.98	65.14

E More Visualization Comparisons

We provide a more comprehensive qualitative comparison in Figure 9, including representative generative and unified baselines such as FLUX [3], Janus-Pro [6], BLIP3-o [5], Show-o2 [66], and BAGEL [9].

F Editing Evaluations

We present a more comprehensive evaluation on image editing on ImgEdit [68] and KRIS-Bench [62], in Table 17 and Table 18 respectively.

Table 17: Quantitative comparisons on ImgEdit.

Model	Bg.	Sty.	Adj.	Ext.	Rem.	Rep.	Add.	Cmp.	Act.	Ovr.
FLUX-Kontext	3.89	4.62	3.69	1.81	2.97	4.20	3.80	3.00	4.18	3.6
BAGEL	3.38	4.53	3.58	1.49	3.15	3.82	3.71	2.64	4.21	3.38
+UNO	4.01	4.83	4.14	2.2	3.58	4.40	4.10	3.06	4.32	**3.85**

Table 18: Quantitative comparisons on KRIS-Bench.

Model	AP	SP	TP	SS	NS	LP	ID	Ovr.
BAGEL-think	67.42	68.33	58.67	63.55	61.40	48.12	50.22	60.18
+UNO	73.09	72.25	59.18	64.65	61.51	48.00	55.67	**62.57**

G Evaluation Robustness

We report the mean and variance of the main evaluation with 4 random seeds in Table 20, demonstrating the statistical robustness of our evaluation.

Table 20: Mean and variance of UNO.

Statistics	GenEval2	DPG-Bench	UniGenBench++
Mean	75.1	86.12	65.14
Variance	0.034	0.024	0.017

H Detailed Prompt List

We illustrate the detailed prompts for generated images shown in Figure 5 in Table 19.

Image	Prompt
Row 1, Col 1	In a surrealist oil painting, the body and horns of a giraffe are composed of bright stars, and many butterflies, also composed of nebulae, are resting on its horns.
Row 1, Col 2	An open book the size of a mountain range, pages crumbling into sandy dunes with ancient scripts glowing, a wanderer climbing the spine amid a sandstorm, warm amber lights and swirling dust particles
Row 1, Col 3	An astronaut is riding on the back of a giant dragon composed of brilliant stardust and diffuse nebulae; the dragon's body is translucent and iridescent. They are shuttling at high speed through Saturn's magnificent rings, with countless glowing ice crystals and rock particles dancing around. In the distant background is Saturn's huge orange-yellow sphere. The entire scene presents a magnificent oil painting texture, with heavy brushstrokes, rich and saturated colors, and strong contrast between light and shadow. The light from the stardust illuminates the astronaut's helmet and deep space, filled with an epic and grand momentum.
Row 1, Col 4	A fantasy city made of huge candies and biscuits, all the buildings and streets in the city present a sweet and seductive texture.
Row 1, Col 5	A crystal clear crystal ball is wrapped in a collapsing Great Wall, surrounded by fine snowflakes.
Row 2, Col 1	Draw a grand piano made of transparent crystal. When the keys are pressed, swarms of glowing butterflies fly out instead of musical notes.
Row 2, Col 2	a miniature house inside an apple and you can see little, cute pandas living inside the apple, eating dinner at the dinner table, rendered hyper-realistic with strong depth of field to make it seem very miniature
Row 2, Col 3	A glowing giant whale whose body is made of a nebula has a coral reef made of broken planets growing on its back.
Row 2, Col 4	A realistic blurred photo of a vibrant city street at golden hour. In the foreground, a glowing neon line art of a joyful boy riding a skateboard. He is outlined in thick glowing white lines, wearing a hoodie with sketchy lightning bolts and stars around him. his shoes glow as he balances with one foot up. Graffiti-style hearts and sparkles trail behind his like motion lines.
Row 2, Col 5	A mesmerizing hyper-realistic digital sketch by artist PJ, featuring an egg cracked open to reveal the miniature, intricate city of London within.The eggshell fragments, both upright and flat, create a chaotic yet harmonious composition against the smooth, white background. The vibrant cityscape, with Tower Bridge, River Thames, Big Ben, Buckingham Palace, and the most charming Londonian landmarks, contrasts beautifully with the muted tones of the kitchen counter in the background. The artist's use of textures and shadows adds depth, while the vibrant colors and gritty style make this scene truly captivating, kitchen background

Table 19: Detailed prompt list for qualitative image generation results in Figure 5.

I Societal Impact and Safeguards

This paper aims at foundational research and not tied to particular applications, let alone deployments. However, we point out that a general improvement in the quality of generative models proposed in this paper could be used to generate Deepfakes for disinformation. The released models should be used in accordance with the Apache-2.0 license and accompanying terms of use, which are intended to mitigate potential abuse.

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 main claim made in the abstract is to convey an important new perspective in unified modeling. Rather than conventional approaches that attempt to simply balance understanding and generation through data manipulation and architectural decoupling, we show that the synergistic effect from understanding to generation can be achieved by explicitly encouraging the gradient flow from understanding to generation, opening a promising new direction toward better unified models achieved through deeper integration. Extensive experiments with two unified model backbones across multiple image generation and editing benchmarks validate our claim.

   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: The paper includes a limitations section in Section 4.3 discussing that the proposed method does not currently improve performance on certain specialized domains.

   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: We provide a detailed list of training hyper-parameters in Section A. Detailed generation training data composition and statistics are provided in Section C and Section B respectively. Training on image editing is conducted on a specified publicly available open-source dataset.

   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: [Yes]

   Justification: We will release the training codes upon acceptance.

   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: We provide a detailed list of training hyper-parameters in Section 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: [Yes]

   Justification: Mean and std of our evaluation with 4 random seeds are provided in Section G.

   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: We provide training resources and compute cost in Section 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: Research conducted in the paper conform to 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: We provide a section discussing the societal impacts in Section I.

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: We discuss safeguards for the released model in Section I.

    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: Our code is developed upon BAGEL [9] and Show-o [66], both under the Apache-2.0 license. Image generation data is collected and filtered from LAION [40], JourneyDB [43] and OpenImages [23] under CC-BY 4.0, custom terms of usage and CC-BY 4.0 license respectively. Image editing data is based on CrispEdit-2M [8] under the CC-BY 4.0 license. All terms of usage under the respective licenses are followed.

    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: [Yes]

    Justification: Codes and models will be released upon acceptance.

    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: [N/A].

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    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: [N/A].

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

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: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    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.