

BLIP3o-NEXT: Next Frontier of Native Image Generation

Jiuhan Chen^{1,2*} Le Xue^{1*} Zhiyang Xu^{3*} Xichen Pan⁴ Shusheng Yang⁴
Can Qin¹ An Yan¹ Honglu Zhou¹ Zeyuan Chen¹ Lifu Huang⁶ Tianyi Zhou²
Junnan Li¹ Silvio Savarese^{1‡} Caiming Xiong^{1‡} Ran Xu^{1‡}

¹Salesforce Research

²University of Maryland ³Virginia Tech ⁴New York University ⁶UC Davis

*Equal Contribution. ‡Corresponding Authors.

Abstract

We present BLIP3o-NEXT, a fully open-source foundation model in the BLIP3 series that advances the next frontier of native image generation. BLIP3o-NEXT unifies text-to-image generation and image editing within a single architecture, demonstrating strong image generation and image editing capabilities. In developing the state-of-the-art native image generation model, we identify four key insights: (1) Most architectural choices yield comparable performance; an architecture can be deemed effective provided it scales efficiently and supports fast inference; (2) The successful application of reinforcement learning can further push the frontier of native image generation; (3) Image editing still remains a challenging task, yet instruction following and the consistency between generated and reference images can be significantly enhanced through post-training and data engine; (4) Data quality and scale continue to be decisive factors that determine the upper bound of model performance. Building upon these insights, BLIP3o-NEXT leverages an Autoregressive + Diffusion architecture in which an autoregressive model first generates discrete image tokens conditioned on multimodal inputs, whose hidden states are then used as conditioning signals for a diffusion model to generate high-fidelity images. This architecture integrates the reasoning strength and instruction following of autoregressive models with the fine-detail rendering ability of diffusion models, achieving a new level of coherence and realism. Extensive evaluations of various text-to-image and image-editing benchmarks show that BLIP3o-NEXT achieves superior performance over existing models.

- 🌐 Website <https://jiuhaichen.github.io/BLIP3o-NEXT.github.io>
- 💻 Code <https://github.com/JiuhaiChen/BLIP3o>

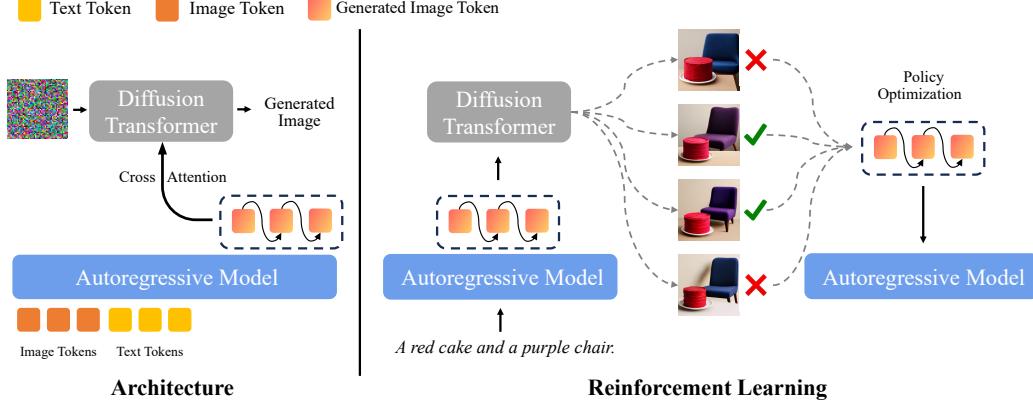


Figure 1: The architecture of BLIP3o-NEXT (left) and its reinforcement learning pipeline (right). BLIP3o-NEXT adopts an Autoregressive (AR) + Diffusion design, where the AR module autoregressively generates image conditions for the diffusion model. The model is jointly optimized with both AR and diffusion objectives. During reinforcement learning, rollouts are rendered from the diffusion transformer, and policy optimization is performed directly on the AR model, enabling seamless integration with existing RL infrastructures originally developed for language models.

1 Introduction

Recently image generation has become a cornerstone of modern artificial intelligence, empowering models to not only synthesize photorealistic images from textual descriptions but also edit existing images with remarkable precision and semantic consistency [2, 26, 6, 40, 10, 39]. These capabilities have redefined how machines interpret, represent, and communicate visual information, enabling a wide range of creative and practical applications, from content creation and design to scientific visualization and simulation.

In this paper, we introduce BLIP3o-NEXT, a novel image generation foundation model that advances the frontier of multimodal learning through innovations in model architecture, multi-task pretraining, reinforcement learning, and comprehensive data engineering. BLIP3o-NEXT adopts a hybrid Autoregressive + Diffusion architecture [26, 6]. The autoregressive model [3] takes a user prompt together with a set of reference images (for image editing) and generates a sequence of discrete image tokens conditioned on the multimodal input. The diffusion model [44] subsequently leverages the final hidden states of generated discrete tokens as conditioning signals to synthesize the final image. This architecture combines the semantic compositionality and global structure understanding captured by the autoregressive model with the fine-grained detail rendering capability of diffusion models, enabling BLIP3o-NEXT to produce images that are both semantically coherent and photorealistically detailed.

The autoregressive model is trained on three primary tasks: (1) text-to-image generation, (2) input image reconstruction, and (3) image editing, which together equip the model with the fundamental capabilities required for diverse downstream tasks. In the post-training stage, beyond fine-tuning on carefully curated high-quality datasets [6, 40, 8], we perform extensive reinforcement learning (RL) to further enhance the model’s text rendering quality and instruction following ability. We propose an efficient RL framework specifically tailored for the BLIP3o-NEXT architecture by leveraging the discrete image tokens. This design allows seamless integration with existing RL infrastructures originally developed for language models, while achieving performance comparable to recent flow-based RL approaches.

Taking advantage of the Autoregressive + Diffusion design, BLIP3o-NEXT exhibits great flexibility in integrating the VAE features [7] of reference images for editing tasks. Through empirical experiments, we find that concatenating VAE features to (1) hidden states generated by the autoregressive model as conditions, and (2) the random noise as initial input in the diffusion process, achieves the highest consistency and visual fidelity. Additionally, powered by the strong multimodal reasoning and

understanding capability of its autoregressive backbone, BLIP3o-NEXT can faithfully follow user instructions for image editing.

In summary, BLIP3o-NEXT makes the following key contributions:

- **A novel and scalable Autoregressive + Diffusion architecture** that advances the next frontier of native image generation.
- **An efficient reinforcement learning method for image generation** that can be seamlessly integrated with existing RL infrastructures for language models, improving text rendering and instruction following abilities.
- **Systematic studies on improving consistency in image editing**, including strategies for integrating VAE features from reference images.
- **Strong performance across diverse benchmarks**, comprehensive evaluation on text-to-image generation benchmarks and image-editing benchmarks reveals that BLIP3o-NEXT consistently outperform existing models.

To support future research and uphold the open-source philosophy of the BLIP3 family, we fully release BLIP3o-NEXT, including pretrained and post-trained model weights, datasets, detailed training and inference code, and evaluation pipelines, ensuring complete reproducibility. We hope that our work will foster continued progress in native image generation.

2 Overview of Architectures for Native Image Generation

2.1 Autoregressive + Diffusion

In recent developments in native image generation, approaches that integrate autoregressive models with diffusion models continue to achieve state-of-the-art performance. In these frameworks, the autoregressive backbone encodes the input prompt, whether text or other modalities, and produces conditions for the diffusion model to generate the final image. Early explorations include Emu2 [34] and Seed-X [11]. Following the success of GPT-4o [2], subsequent works such as MetaQuery [26], UniGen [36], and BLIP3-o [6] push the direction further, while recent models like Qwen-Image [39], Gemini Nano Banana and Seedream [29] continue to push the performance frontier. The advantage of Autoregressive + Diffusion is its **simplicity** and **scalability**, and also leveraging the strengths of existing autoregressive models, such as transferring instruction following and reasoning capabilities into image generation procedure.

Within this framework, the question is how to derive the conditioning signal from the autoregressive model for the diffusion model. Existing approaches can be broadly grouped into the following categories: (1) The autoregressive model directly generates continuous embeddings (e.g., CLIP representations), which are then used as conditions for the diffusion model, similar to Emu2 [34] and MetaMorph [37]. (2) The autoregressive model compresses all input information, including text and images, into a fixed number of learnable query, as demonstrated by Seed-X [11], MetaQuery [26], and BLIP3-o [6]. (3) The autoregressive model encodes both image and text inputs, and the hidden states from the autoregressive model are directly used as the conditioning signal for the diffusion model, as in OmniGen2 [40], Qwen-Image [39], and MANZANO [20].

Since continuous embeddings and learnable query tokens generate conditioning signals by sampling from AR models, they tend to perform better on image generation tasks requiring reasoning capability compared with directly using hidden states [35]. However, both continuous embeddings and learnable query compress all conditioning information into a fixed number of tokens, which inevitably limits their representational capacity. For example, Emu2 [34] and BLIP3-o [6] usually use 64 tokens. In contrast, using hidden states offers a more flexible and efficient way for encoding the condition.

2.2 Mixture of Transformers

LMFusion [32] and BAGEL [10] adopt a different architecture for native image generation. At a high level, they employ two transformer experts to separately process understanding and generation information, while tokens from different modalities interact through shared multimodal self-attention within each transformer block. Although this structure facilitates richer information sharing across modalities, it faces challenges in flexibility and scalability, and suffers from high inference latency.

2.3 BLIP3o-NEXT Architecture

In BLIP3o-NEXT, we still adopt an Autoregressive + Diffusion architecture, where an autoregressive model first predicts discrete image tokens from multimodal inputs, and the hidden states of these tokens are subsequently used to condition a diffusion model for high-fidelity image synthesis. Unlike previous approaches, where the autoregressive model takes in and generates continuous image embeddings [11, 37, 26, 6], our autoregressive model operates on discrete visual tokens. Specifically, each image is first encoded using the SigLIP2 model [13], and the resulting continuous embeddings are quantized into a finite vocabulary of tokens, yielding 729 discrete tokens per image given the image resized to 384x384. Conditioned on a text prompt (or reference images for image editing), the autoregressive model is trained through next-token prediction over discrete image tokens. The diffusion model is then applied on top of the autoregressive outputs to refine fine-grained details. Specifically, the hidden states of the predicted tokens serve as conditioning signals for the diffusion model to diffuse the final VAE features, which produce the final high-fidelity images, as illustrated in Figure 1 (left).

During training, BLIP3o-NEXT is optimized with two objectives:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{diff}}, \quad (1)$$

where \mathcal{L}_{CE} denotes the cross-entropy loss over text and discrete image tokens in autoregressive model, $\mathcal{L}_{\text{diff}}$ is the diffusion loss, and λ is a balancing weight. Due to computational constraints, we initialize the autoregressive model with Qwen3 [47] and the diffusion transformer with SANA1.5 [44], resulting in a total of approximately 3B parameters. For training, we employ BLIP3o-Pretrain [6] as the pretraining corpus and BLIP3o-60K [6] as the instruction tuning dataset.

2.4 Discussion

From GPT-4o, where an autoregressive model produces discrete image tokens and is refined by a diffusion model, to more recent models such as Qwen-Image [39], architectures are increasingly dominated by diffusion backbones. For example, Qwen-Image integrates a 7B vision language model with a 20B diffusion transformer. In practice, GPT-4o exhibits higher inference latency, whereas advances in accelerating diffusion transformers [30, 28, 22] have enabled diffusion-centric models to achieve both greater scalability and faster inference efficiency.

Under Autoregressive + Diffusion framework, we observe that most architectures deliver comparable performance, with minor design variations yielding marginal differences. Therefore, an architecture can be considered effective, as long as it remains simple, scalable, and supports fast inference.

3 Image Generation with Reinforcement Learning

Reinforcement learning (RL) has been extensively explored in large language models, but remains relatively underexplored in the image generation domain. GPT-4o [2] is the representative native image generation model to successfully integrate RL. In this work, we primarily discuss two directions for applying RL to image generation: (1) applying RL to the autoregressive mode under the BLIP3o-NEXT framework, and (2) applying RL to the diffusion model, for example Flow-GRPO[23].

3.1 RL for Autoregressive Model

Applying RL to the autoregressive model is a natural extension of language modeling. The key advantage of this approach is its compatibility with existing RL frameworks developed for language models, allowing most training and inference infrastructures to be easily adapted for image generation training.

We use the Group Relative Policy Optimization (GRPO) algorithm [31] under BLIP3o-NEXT framework. For each text prompt $p \sim \mathcal{D}$, the previous policy $\pi_{\theta_{old}}$, i.e., the autoregressive model, samples a group of G trajectories $\{o_1, \dots, o_G\}$. Each of the trajectory consists of 729 discrete image tokens for an image. These trajectories are then decoded by a frozen diffusion model to generate the corresponding images $\{I_1, I_2, \dots, I_G\}$, each assigned a reward score $\{r_1, \dots, r_G\}$ by the reward

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color	Attri.	Overall
FLUX.1-dev (12B) [18]	0.98	0.93	0.75	0.93	0.68	0.65	0.82	
OmniGen2 (7B) [40]	1.00	0.95	0.64	0.88	0.55	0.76	0.80	
Qwen-Image (27B) [39]	0.99	0.92	0.89	0.88	0.76	0.77	0.87	
Metaqueries XL (7B) [26]	—	—	—	—	—	—	—	0.80 [†]
BAGEL (14B) [10]	0.98	0.95	0.84	0.95	0.78	0.77	0.88 [†]	
BLIP3o (8B) [6]	—	—	—	—	—	—	—	0.84
BLIP3o-NEXT-GRPO-GenEval (3B)	0.99	0.95	0.88	0.90	0.92	0.79	0.91	

Table 1: Quantitative results on GenEval benchmarks. ([†] denotes the rewritten prompts.)

function. The rewards are normalized across the group to compute the advantages A_i , following the GRPO procedure. During training, the diffusion model remains frozen, while the policy model π_θ is optimized by maximizing the following objective:

$$\begin{aligned} \mathcal{L}_{GRPO}(\theta) = & \mathbb{E}_{p \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot | p)} \\ & \frac{1}{G} \sum_{i=1}^G \left(\min\left(\frac{\pi_\theta(o_i | p)}{\pi_{\theta_{old}}(o_i | p)}, A_i\right), \text{clip}\left(\frac{\pi_\theta(o_i | p)}{\pi_{\theta_{old}}(o_i | p)}, 1 - \epsilon, 1 + \epsilon\right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_\theta \| \pi_{\theta_{ref}}) \end{aligned} \quad (2)$$

Here, ϵ and β are hyperparameters, $\pi_{\theta_{ref}}$ is a fixed reference policy, and $\mathbb{D}_{KL}(\pi_\theta \| \pi_{\theta_{ref}})$ is a KL divergence penalty that constrains policy updates relative to the reference policy. RL training can be seen in Figure 1 right.

3.2 RL for Diffusion Model

Instead of autoregressive model, RL can also be applied to the diffusion model. Qwen-Image [39] is the representative model to successfully apply both DPO [27] and Flow-GRPO [23] within a native image generation model.

After the DPO stage, Qwen-Image uses additional fine-grained reinforcement learning with GRPO, following the Flow-GRPO framework [23]. Conditioned on the hidden state of the text prompt p from the autoregressive module, the flow model generates a set of G candidate images $\{x_0^i\}_{i=1}^G$ along with their trajectories $\{x_T^i, x_{T-1}^i, \dots, x_0^i\}_{i=1}^G$. The training objective of Flow-GRPO is consistent with Equation 2.

A key distinction from RL in autoregressive models is trajectory sampling. In Flow-GRPO, trajectories $\{x_T^i, \dots, x_0^i\}_{i=1}^G \sim \pi_\theta$ are sampled according to the flow matching dynamics:

$$dx_t = v_t dt, \quad (3)$$

where $v_t = v_\theta(x_t, t, p)$ is the velocity predicted by the model. However, this deterministic formulation lacks stochasticity and thus fails to support adequate exploration. To address this, Flow-GRPO reformulates trajectory sampling as a stochastic differential equation (SDE), injecting randomness into the process:

$$dx_t = \left(v_t + \frac{\sigma_t^2}{2t} (x_t + (1-t)v_t) \right) dt + \sigma_t dw_t, \quad (4)$$

where dw_t denotes Brownian motion and σ_t controls the magnitude of randomness.

In addition to Flow-GRPO, DanceGRPO [46] extends GRPO to a broader range of visual generation tasks, including text-to-image, text-to-video, and image-to-video. It reformulates both diffusion sampling and rectified flows within the framework of stochastic differential equations, enabling GRPO to generalize across different architectures and training paradigms. MixGRPO [19] further improves computational efficiency by combining SDE and ODE based formulations.

3.3 Reward Models

Reward functions can be broadly divided into two categories. (1) Verifiable rewards, such as GenEval [12] for the composition of multiple objects and OCR-based evaluation for visual text

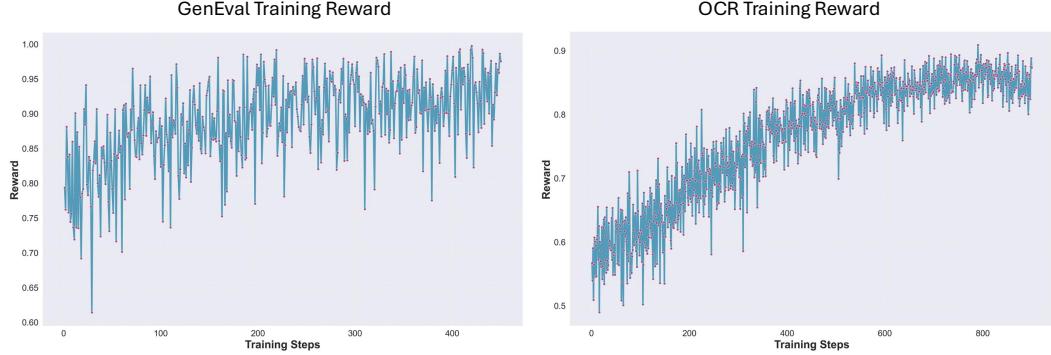


Figure 2: Training reward for GenEval and visual text rendering.

rendering. (2) Model-based rewards, including PickScore [17], ClipScore [14], HPSv2.1 [42], ImageReward [45], and UnifiedReward [38], which assess image quality, image–text alignment, and human preference.

3.4 Experiments

In our experiments, we only apply RL for autoregressive model and focus on two verifiable reward tasks. (1) Multiple object composition: training prompts are adopted from Flow-GRPO [23], and evaluation is performed with the GenEval evaluator. (2) Visual text rendering: training prompts are also sourced from Flow-GRPO [23], and evaluation relies on PaddleOCR [9]. Figure 2 clearly illustrates the increasing reward trend throughout training. Figures 3 and 4 present qualitative comparisons before and after GRPO training, demonstrating noticeable improvements in both object composition and text rendering quality.

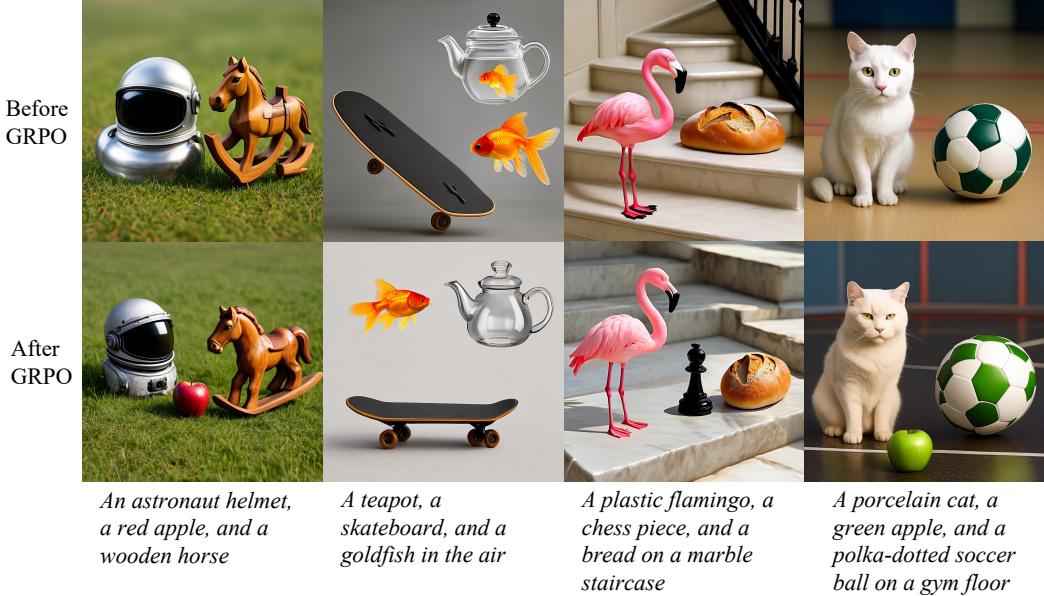


Figure 3: Qualitative results of multiple object composition before and after GRPO.

3.5 Discussion

The model architecture determines whether RL is applied to the autoregressive or diffusion model. For example, in Qwen-Image, the autoregressive module mainly serves as a text encoder, while the diffusion model performs most of the image generation. RL is more effective when applied to the diffusion model for Qwen-Image. In contrast, within the BLIP3o-NEXT framework, the

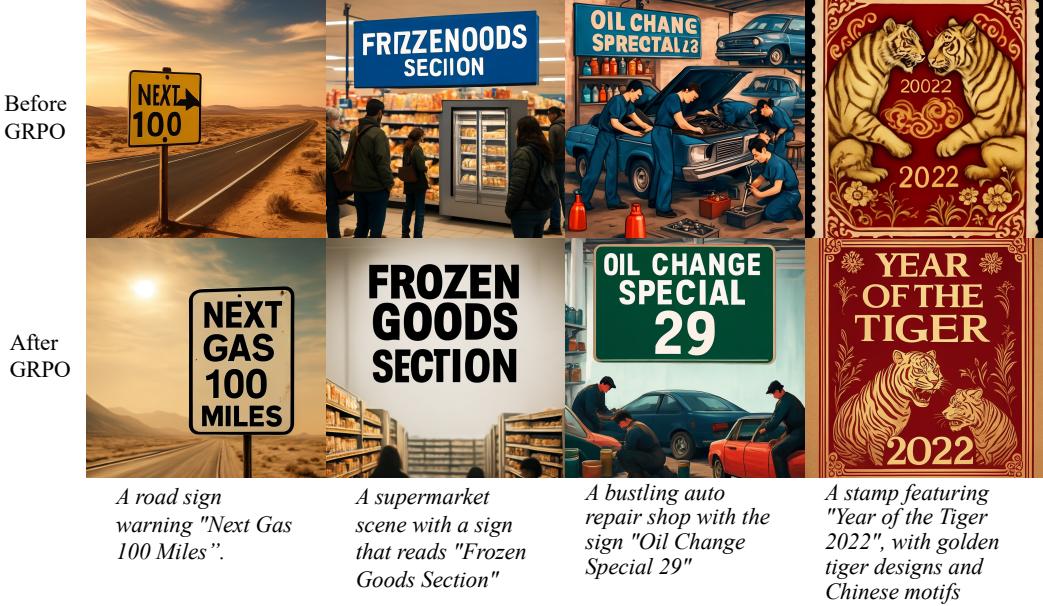


Figure 4: Qualitative results of text rendering before and after GRPO.

autoregressive model is responsible for producing image tokens, and the diffusion model primarily functions as an image decoder. The autoregressive model plays a central role in image generation, making it the natural target for RL optimization. Empirically, without diffusion acceleration, applying RL to diffusion model is slower due to the lack of KV cache support and the need for multiple time steps.

The central challenge in applying reinforcement learning to native image generation lies in metric design rather than the RL algorithm itself. The key open problem is to develop reward models that can effectively capture and balance multiple dimensions, including image quality, instruction following, and human preference alignment.

4 Image Editing

To incorporate reference images as multimodal conditions for image generation, a common strategy is to feed them into the autoregressive backbone. Recent models such as MetaQuery [26], BLIP3-o [6], OmniGen2 [40] and Qwen-Image [39] are built on top of existing vision–language models such as Qwen-VL [3], which natively support image inputs as the conditions. Within the BLIP3o-NEXT framework, this is achieved by converting the reference image into quantized image tokens and concatenating them with the input text prompt tokens.

The primary challenge in image editing is maintaining consistency between the generated and reference images. To address this issue, we adopt the following strategies to enhance consistency.

Image Reconstruction Task. We perform an image reconstruction task, where the reference image is provided as input and the text prompt is “*Keep the image unchanged.*”. This task encourages the model to faithfully reconstruct visual details and align the generative process with the conditioning image.

Conditioning on VAE Latents. While the reference image is represented through a semantic vision encoder to autoregressive model, such representations often lack fine-grained pixel-level information. To address this limitation, we incorporate low-level, detail-preserving VAE latents as additional conditioning signals for the diffusion model. Specifically, we explore two complementary methods for integrating these VAE latents: (1) cross-attention conditioning and (2) noise-space injection, both

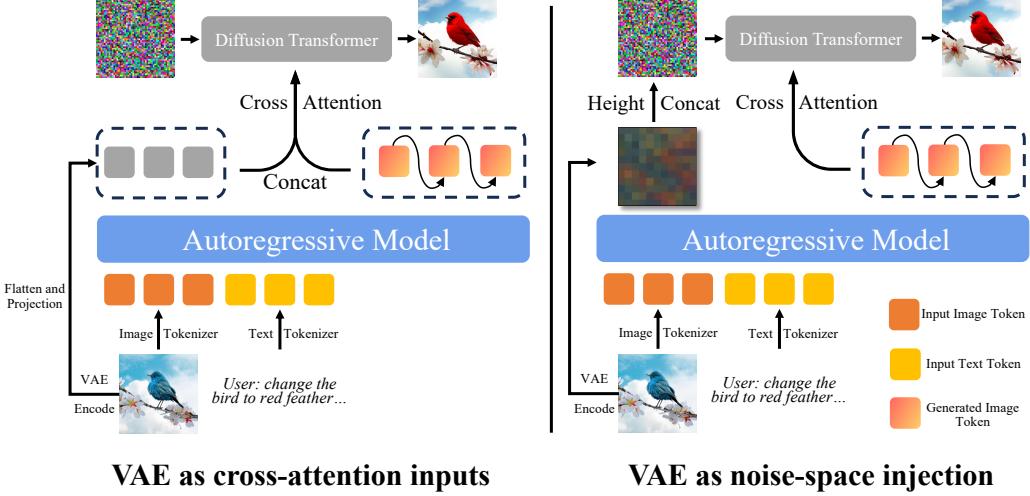


Figure 5: Comparison of VAE feature integration strategies in BLIP3o-NEXT. In the *VAE as cross-attention inputs* setup, the flattened VAE tokens are appended to the multimodal tokens produced by the autoregressive model. Empirically, we find that combining both methods yields the best visual consistency.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall ↑
MagicBrush [50]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix [4]	2.45	1.83	1.44	2.07	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [49]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
OmniGen [43]	3.67	3.39	1.71	2.94	2.43	3.41	4.19	2.24	3.38	2.93
ICEdit [51]	3.39	3.39	1.73	3.15	2.93	3.08	3.62	2.09	3.06	2.95
StepX-Edit [24]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
BAGEL [10]	3.76	3.04	1.70	3.43	3.04	3.40	4.64	2.64	2.62	3.25
UniWorld-V1 [21]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [40]	3.57	3.06	1.77	3.74	3.02	3.57	4.81	2.52	4.68	3.44
FLUX.1 Kontext (Pro) [18]	4.15	2.35	4.56	3.57	3.42	4.56	4.57	3.63	4.63	4.00
GPT Image 1 [2]	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
Qwen-Image [39]	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
BLIP3o-NEXT (3B)	4.00	3.78	2.39	4.05	2.61	4.30	4.64	2.67	4.13	3.62

Table 2: Image editing benchmark results for ImgEdit [48], with all metrics evaluated using GPT-4.1. The “Overall” score is computed as the average across all task categories. While our 3B model still lags behind GPT-Image and Qwen-Image, it achieves comparable performance to BAGEL and OmniGen2.

designed to enhance visual consistency with respect to the reference image. Empirically, we find that combining both approaches yields the best consistency.

(1) VAE features as cross-attention inputs in DiT. After extracting the reference image’s VAE features, we flatten the spatial dimensions (height and width) and project the channel dimension to match the cross-attention input size of the DiT. The resulting projected features are then concatenated with the multimodal context tokens produced by the autoregressive model. These concatenated tokens serve as the input to the DiT’s cross-attention layers, enabling the model to incorporate both semantic and low-level cues, as shown in Figure 5 left.

(2) VAE features as noise-space injection. Alternatively, we concatenate the extracted VAE features with the diffusion noise tensor along the height dimension, effectively augmenting the noise input with adjacent reference image VAE features. This composite input is then fed into the DiT during denoising. After the model predicts the refined noise, we compute the loss only over the region corresponding to the original random noise, using it as the flow-matching loss objective as shown in Figure 5 right.

4.1 Experiments

In our image editing experiments, we adopt a multitask learning setup that jointly trains the model on both image reconstruction and image editing objectives. The training data is curated from various open-source datasets, including BLIP3-o [6], OmniGen2 [40], ShareGPT4V [8], and Awesome-Nano-Banana [1]. To increase the data scale and stabilize training, we repeat selected subsets to construct a larger dataset ensemble. The final training corpus contains approximately 10 million samples, including repeated entries. Table 2 presents the results on the ImgEdit [48] benchmark, where our 3B model achieves performance comparable to larger models such as BAGEL and OmniGen2. Figure 6 illustrates qualitative results on consistency. The comparison between models with and without VAE latents demonstrates that incorporating VAE latents effectively enhances consistency. However, since the VAE in SANA uses a downsampling ratio of 32 to accelerate training and inference [44], the generated images still exhibit slight inconsistencies with the reference images.

4.2 Future Exploration

Besides image reconstruction and conditioning on VAE latents, we also want to highlight the following strategies:

Reinforcement Learning for Image Editing. Applying RL to image editing offers a promising direction for improving instruction following and consistency between the generated output and the reference input. While model-based reward functions can effectively guide instruction following [41], the design of reward models for measuring consistency remains relatively underexplored.

Designing System Prompts for Inpainting and Subject-Driven Generation. Another avenue of exploration is the design of system prompts that explicitly distinguishes between inpainting and subject-driven generation tasks. Inpainting tasks prioritize spatial and background consistency, whereas subject-driven generation emphasizes maintaining consistency in the subject’s appearance. Therefore, incorporating task-specific system prompts can effectively guide the model to handle these two categories of editing more appropriately.

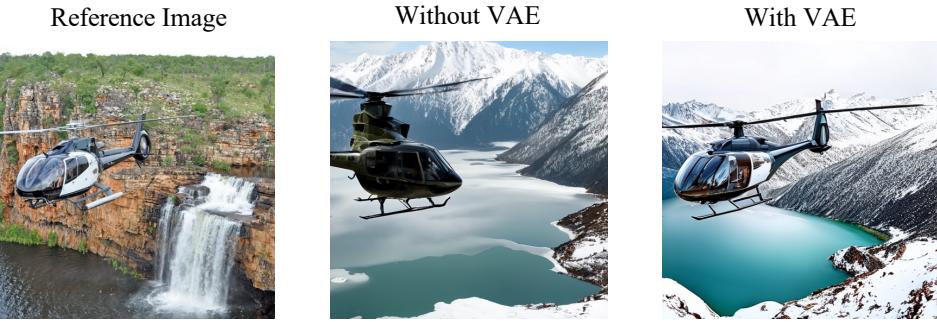
Prompt Rewriting. In practice, we use prompt rewriting to enrich prompt details and improve instruction following abilities for both image generation and editing tasks.

While Autoregressive + Diffusion architectures enable strong instruction following, editing consistency still lags behind, even with VAE feature injected into the diffusion model. Bridging this gap calls for advances in data engineering and scaling reinforcement learning specifically tailored for image editing.

5 Training Recipe and Evaluation

Data quality remains a decisive factor in determining the overall performance of the model. Although different models have different data engine pipelines, they generally share several key components. (1) Diversity: We categorize image topics into domains such as Environments, Business, Cities, Food & Drink, Nature, Objects, Pets, Wildlife, and Lifestyle. The dataset integrates publicly available sources, including CC12M [5], SA-1B [16], and JourneyDB [33], supplemented with additional proprietary images. (2) Filtering: We perform extensive data cleaning; for example, we remove extremely low-resolution or corrupted images, and we exclude samples containing watermarks, etc. (3) Captioning: Dense captions are generated using Qwen-VL-2.5 [3], samples with overly long captions (exceeding 120 tokens) or low CLIP-based image-text alignment scores are discarded. (4) Synthetic data: To enrich training diversity, we further construct synthetic datasets, particularly for text-rendering tasks, and distill data from frontier models.

Regarding the evaluation, although there are numerous benchmarks for evaluating image generation performance [12, 15, 25], there remains a significant need for more specialized benchmarks, particularly for image editing. Such benchmarks are essential for assessing a model’s instruction following ability and the consistency between the generated images and the reference inputs.



Change the waterfall and rocky cliff background to snowy mountains and icy landscape.



Add a wooden canoe drifting along the river in the foreground.



Change the green field and trees in the background to a snowy landscape.



Add a person wearing a red winter coat and black snow pants walking across the snowy field near the center of the image.

Figure 6: Qualitative results for image editing comparing models with and without VAE latent condition.

6 Conclusion

In this work, we introduced BLIP3o-NEXT, a fully open-source foundation model that advances the frontier of native image generation by unifying text-to-image synthesis and image editing within a single architecture. Through extensive exploration, we identify four central insights: architectural

simplicity coupled with scalability is often sufficient; reinforcement learning holds strong potential to further improve generation quality; post-training remains crucial for instruction following and editing consistency; and high-quality, large-scale data continues to set the performance ceiling. Building on these principles, BLIP3o-NEXT integrates the instruction following and reasoning capabilities of autoregressive modeling with the fine-grained rendering ability of diffusion, yielding coherent and high-fidelity visual outputs. Experimental results across diverse benchmarks demonstrate that BLIP3o-NEXT achieves superior performance in both generation and editing tasks, confirming the effectiveness of the Autoregressive + Diffusion paradigm.

Looking forward, we believe the findings presented here point toward promising directions for the next frontier of foundation models, where unified architectures, reinforcement learning, and scalable post-training jointly drive progress in controllable, instruction-aligned, and high-quality native image generation systems.

References

- [1] awesome-nano-banana. <https://github.com/JimmyLv/awesome-nano-banana>, 2025.
- [2] Introducing 4o image generation. <https://openai.com/index/introducing-4o-image-generation/>, 2025.
- [3] 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.
- [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] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.
- [6] Juhai 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.
- [7] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.
- [8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023.
- [9] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025.
- [14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- [15] 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.
- [16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

- [17] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.
- [18] 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.
- [19] Junzhe Li, Yutao Cui, Tao Huang, Yingping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.
- [20] Yanghao Li, Rui Qian, Bowen Pan, Haotian Zhang, Haoshuo Huang, Bowen Zhang, Jialing Tong, Haoxuan You, Xianzhi Du, Zhe Gan, et al. Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer. *arXiv preprint arXiv:2509.16197*, 2025.
- [21] 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.
- [22] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025.
- [23] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.
- [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] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.
- [26] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Juhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.
- [27] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- [28] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *Advances in Neural Information Processing Systems*, 37:117340–117362, 2024.
- [29] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.
- [30] Huiyang Shao, Xin Xia, Yuhong Yang, Yuxi Ren, Xing Wang, and Xuefeng Xiao. Rayflow: Instance-aware diffusion acceleration via adaptive flow trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18113–18123, 2025.
- [31] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Huawei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [32] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

- [33] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [34] 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.
- [35] Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025.
- [36] Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025.
- [37] 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.
- [38] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.
- [39] 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.
- [40] 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.
- [41] Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhua Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025.
- [42] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.
- [43] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuteng Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.
- [44] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer, 2024.
- [45] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.
- [46] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.
- [47] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 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] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025.
- [50] Kai Zhang, Lingbo Mo, Wenhui 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.
- [51] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.