Nexus-Gen: Unified Image Understanding, Generation, and Editing via Prefilled Autoregression in Shared Embedding Space

Hong Zhang[1,2], Zhongjie Duan[2], Xingjun Wang[2], Yuze Zhao[2], Weiyi Lu [3], Zhipeng Di [3], Yixuan Xu [3], Yingda Chen[2], Yu Zhang[1†]

[1]College of Control Science and Engineering, Zhejiang University
[2]ModelScope Team, Alibaba Group Inc. [3]AIOS Team, Alibaba Group Inc.
{hongzhang99, zhangyu80}@zju.edu.cn
{duanzhongjie.dzj, xingjun.wxj, yuze.zyz, weiyi.lwy, dizhipeng.dzp, xuyixuan.xyx, yingda.chen }@alibaba-inc.com

Abstract

Unified multimodal generative models aim to integrate image understanding and generation abilities, offering significant advantages in harnessing multimodal corpora, particularly interleaved text-image data. However, existing unified models exhibit limitations in image synthesis quality, autoregressive error accumulation, and image editing capability. In this work, we propose Nexus-Gen, a novel architecture that unifies image understanding, generation, and editing tasks in a shared image embedding space. This shared space serves as a bridge for the autoregressive and diffusion models, which seamlessly integrates their complementary strengths in cross-modal modeling. To mitigate the severe error accumulation during autoregressive embedding prediction, we propose a novel prefilled autoregression strategy that aligns training-inference dynamics by prefilling input sequences with learnable embeddings. After multi-stage and multi-task training on our constructed large-scale dataset with 26.3 million samples, Nexus-Gen achieves state-of-the-art performance on the evaluation benchmarks spanning image understanding, generation and editing tasks. All models, datasets, and source codes are released in https://github.com/modelscope/Nexus-Gen to facilitate further advancements across the field.

Introduction

Multimodal generative modeling has emerged as a pivotal frontier in AI research. Multimodal large language models (MLLMs) demonstrate notable competence in image understanding, while diffusion models lead in image generation. To further harness the potential of vision-language modeling, particularly the synergistic benefits across modalities, recent studies focused on MLLMs with unified modality architectures. The core advantage of unified modeling lies in the efficient data utilization and cross-modal representation learning, enabling the integration of multiple cross-modal capabilities. Such integration proves effective for complex tasks like image editing, visual reasoning, and reasoning-enhanced image generation, positioning unified MLLMs as a crucial pathway toward unified general intelligence.

Integrating image understanding and generation within a unified framework while enabling joint optimization remains a fundamental challenge for unified MLLMs. Prior works, including Chameleon (Team 2024), Janus-Pro (Chen et al. 2025b), and Emu3 (Wang et al. 2024), predominantly adopt autoregressive models coupled with variational autoencoders (VAEs) (Kingma, Welling et al. 2013). These frameworks employ autoregressive models to predict visual embeddings, which are fed into VQ-VAE (Sun et al. 2024a) or VAE decoders for image generation. However, they underperform state-of-the-art diffusion models (Podell et al. 2024; Esser et al. 2024; Labs 2024) in image synthesis. This gap is attributed to the lack of pixel-level image modeling capabilities of autoregressive models. Another category of methods, such as SEED-X (Ge et al. 2024) and MetaMerph (Tong et al. 2024b), predicts image embeddings autoregressively and employs diffusion models for image generation. A key limitation of these works is the unresolved error accumulation phenomenon during autoregressive token-by-token generation of continuous embeddings. Additionally, the extensibility of these frameworks to downstream tasks (e.g., image editing) remains insufficiently validated.

To leverage the rich pretrained knowledge of autoregressive LLMs and diffusion models, this paper proposes Nexus-Gen, a framework that unifies image understanding, generation, and editing tasks within a shared continuous image embedding space. As illustrated in Figure 1(a), this embedding space serves as a pivotal interface between the autoregressive model and diffusion vision decoder. It also plays a vital role in preserving information integrity while maintaining embedding versatility across diverse tasks. For image understanding task, input images are encoded into the unified space to predict textual outputs. For image generation and editing tasks, the autoregressive model generates the target image embeddings within the space, which are subsequently decoded into images by vision decoders. Crucially, we uncover that autoregressive prediction of target image embeddings suffers from error severe accumulation. To address this, we propose the prefilled autoregression strategy, which prefills the input sequence with several learnable embeddings to align training and inference dynamics.

To perform joint optimization across multiple tasks and fully utilize the multimodal corpus, we construct a large-scale dataset of 26.3 million samples and propose a multi-stage training strategy for Nexus-Gen. The first stage conducts multi-task pretraining of the autoregressive model across image understanding, generation, and editing tasks. This process develops unified any-to-any modal prediction capabilities. The second stage adapts the generation vision

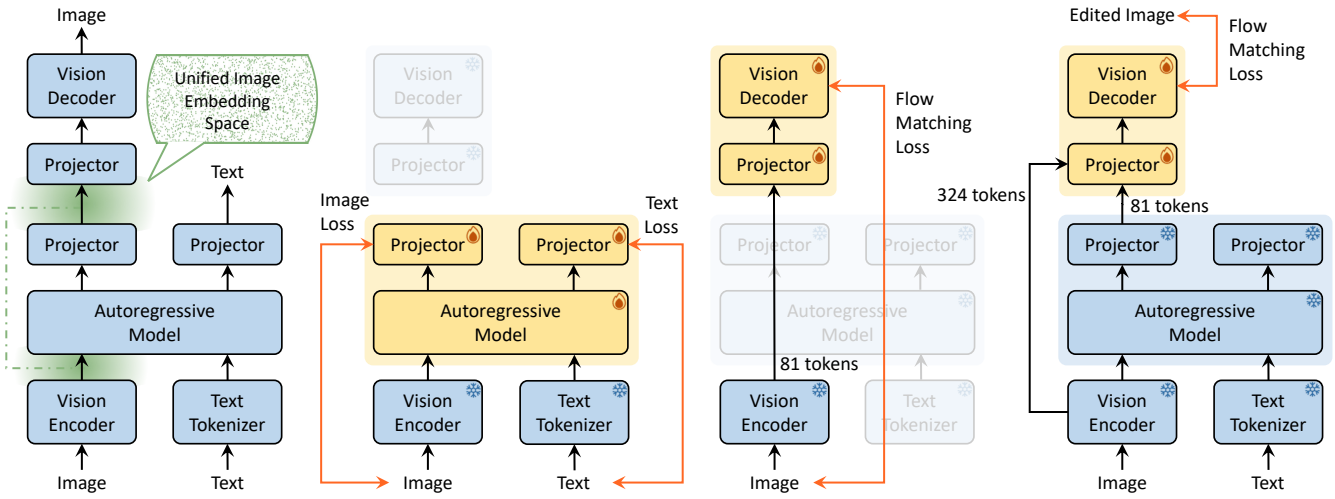


Figure 1: The architecture and the multi-stage training recipe for Nexus-Gen.

decoder by fine-tuning it on a high-quality image generation dataset via replacing the textual conditioning with our unified image embeddings. The third stage adapts the editing decoder by fine-tuning it on our curated high-quality ImagePulse dataset to enable dual-stream image embedding inputs. Through multi-stage training, Nexus-Gen achieves superior performance on image understanding, generation and editing tasks. Specifically, it attains scores of 45.7 on the MMMU understanding benchmark (Yue et al. 2024) and 0.81 on GenEval generation benchmark (Ghosh, Hajishirzi, and Schmidt 2023). Our contributions are as follows:

- We propose Nexus-Gen, a unified model that leverages a unified image embedding space to bridge the capabilities of LLMs and diffusion models.
- We proposed a prefilling strategy that effectively avoids error accumulation during the prediction process of autoregressive models, thereby extending the generative capabilities of autoregressive models.
- We demonstrate the capabilities of Nexus-Gen through extensive experiments. Multiple benchmarks show that Nexus-Gen achieves state-of-the-art performance in image understanding, generation, and editing.
- We curate and release a dataset comprising 26.3 million samples for unified image understanding, generation, and editing tasks to promote research advances.

Related Works

Recent advances in unified architectures for image understanding and generation have stimulated research efforts, leading to two dominant paradigms: autoregressive model with VAE and autoregressive model with diffusion models.

Autoregressive Model with VAE These methodologies exclusively employ lightweight VAEs (Kingma, Welling et al. 2013) or VQ-VAEs (Sun et al. 2024a) for image decoding, positioning the autoregressive model to modeling image information within the pixel space. Notably, Chameleon (Team 2024), Show-O (Xie et al. 2024), and Emu3 (Wang et al. 2024) utilize a VQ-Tokenizer as the visual decoder, training the LLM on interleaved text-image data for unified modeling. Janus (Wu et al. 2025) and Janus-Pro (Chen et al. 2025b) further refine this architecture by decoupling understanding and generation tasks within different encoding spaces. They employ SigLIP (Zhai et al. 2023) for understanding-oriented encoding and VQ-Tokenizer for generative encoding, respectively. Crucially, all aforementioned methods rely on autoregressive prediction of subsequent visual tokens. Conversely, Transfusion (Zhou et al. 2025), and Janus Flow (Ma et al. 2025) adopt diffusion loss to optimize visual token generation.

Autoregressive Model with Diffusion Leveraging pretrained diffusion models as an additional component to synthesize image, these approaches typically yield superior image quality compared to autoregressive model with VAE techniques. The closed-source GPT-4o (OpenAI 2025) model exemplifies this architectural paradigm by adopting the workflow: token $\rightarrow$ [transformer] $\rightarrow$ [diffusion] $\rightarrow$ pixels. Representative open-source frameworks, including SEED-X (Ge et al. 2024), Emu2 (Sun et al. 2024b), and MetaMerph (Tong et al. 2024b), adopt SDXL (Podell et al. 2024) as the vision decoder while utilizing regression loss to optimize the language LLM's visual prediction capability. Regarding architectural variations under this paradigm, LM-Fusion (Shi et al. 2024) and MetaQuery (Zhou et al. 2025) both maintain the LLM in a frozen state. The former trains the vision decoder via shared attention mechanisms, while the latter employs learnable queries as an intermediary bridge between the LLM and the diffusion model. In contrast to the aforementioned methods, this work employs a unified embedding space to jointly model image under-

standing, generation and editing tasks. This design enables the LLM to capture cross-task correlations, facilitating subsequent research into interleaved tasks and reasoning-based multimodal understanding and generation. Additionally, we propose the prefilled autoregression strategy to optimize the generation of continuous image embeddings.

Approach

Architecture

The architecture of Nexus-Gen incorporates three core components, which are depicted in Figure 1(a). The vision encoder and decoder are responsible for unified image embedding, while the autoregressive model facilitates unified multimodal context-aware reasoning.

Unified Image Embedding Space Existing research (Gu et al. 2025) on multimodal representation models demonstrates that unified embedding training across multiple downstream tasks facilitates a more comprehensive understanding of content information than single-task training. Thus, we propose a unified image embedding space to jointly model image-related tasks. For image understanding tasks, images are first encoded into image embeddings via a vision encoder, which are then further interpreted by the autoregressive model. For image generation task, the autoregressive model generates image embeddings based on textual descriptions, and the embeddings are subsequently decoded into images by the vision decoder. By integrating both understanding and generation capabilities, our framework enables image editing ability through modifications to the image embeddings. As unified models evolve towards reasoning-intensive and multi-turn conversational paradigms, our unified embedding space allows for the direct reuse of model-generated embeddings in subsequent reasoning or multi-turn conversations.

Vision Encoder We adopt the vision transformer of Qwen2.5-VL-7B-Instruct (Bai et al. 2025) as our vision encoder and utilize its vision embedding space as our unified image embedding space. This space is implicitly aligned with textual representation space due to the multimodal abilities of the base model, which facilitates the establishment of robust text-image mapping with minimal training data. Leveraging the dynamic resolution capability of the vision encoder, we can modulate the number of image embedding tokens (N_E) by adjusting the input resolution ($H \times W$). This relationship is formally expressed as:

$$N_E = \left\lfloor \frac{H}{P} \right\rfloor \times \left\lfloor \frac{W}{P} \right\rfloor \tag{1}$$

where $P = 14$ denotes the size of each patch. Higher resolutions produce more token embeddings, with each token corresponding to a smaller spatial region and capturing more low-level features. Conversely, lower resolutions yield fewer tokens where each represents a larger spatial region, resulting in embeddings that encode more high-level features.

Autoregressive Model The autoregressive model is utilized for unified multimodal reasoning, as illustrated in Figure 1(a). Textual inputs are tokenized by the text tokenizer and projected into text embeddings, while visual inputs are encoded into image embeddings through the vision encoder. These text and image embeddings are jointly fed into the autoregressive model to predict the embeddings of output tokens. For image synthesis tasks, the output image embeddings predicted by the model are mapped to the unified image embedding space via a vision projector. For text generation, the output text embeddings are converted into logits by the text projector. In Nexus-Gen, the trainable parameters of the autoregressive model and text projector are initialized from Qwen2.5-VL-7B-Instruct (Bai et al. 2025) to inherit the pretrained linguistic abilities. The vision projector is a randomly initialized linear layer for embedding alignment.

To prevent information loss, input images for understanding and editing tasks are encoded to embeddings at their native resolution without downsampling. For output images in generation and editing tasks, there exists a trade-off. Employing more image tokens helps capture finer image details. However, an excessive token count makes the task substantially more difficult for the autoregressive model, leading to degraded generation performance. Through experimental validation, we opted for a token count of 81, ensuring the reconstruction quality of the vision decoder without imposing excessive generation pressure on the autoregressive model.

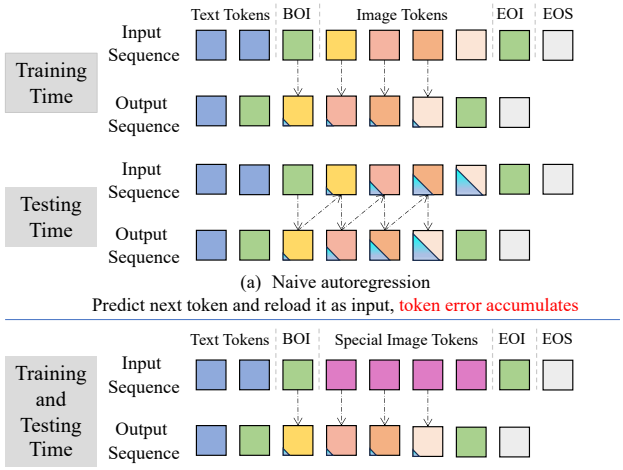
Vision Decoder To achieve high-fidelity image decoding from model-generated embeddings, we adopt the diffusion transformer of FLUX.1-Dev (Labs 2024) as our vision decoder by replacing its native T5 text embeddings (Raffel et al. 2020) with our designed conditioning mechanisms. Given the divergent objectives of image generation and editing tasks, we implement specialized conditioning schemes and architectural configurations for the respective decoders.

For the image generation task, which emphasizes semantic consistency with textual descriptions, the decoder is designed to reconstruct images that are semantically consistent with the 81-token image embeddings produced by the autoregressive model, as presented in Figure 1(c). To condition the diffusion transformer on these embeddings, a two-layer MLP projector is utilized for embedding alignment.

For image editing tasks requiring faithful execution of editing instructions while preserving details in unaltered regions, we propose an editing decoder with dual conditioning mechanisms, as detailed in Figures 1(d). The first condition incorporates 81-token embeddings from the autoregressive model, conveying semantic information of the target image. The second condition integrates 324-token embeddings derived from direct encoding of the input image, preserving fine-grained details of the original content. To effectively model the hierarchical relationships between these two conditions, we introduce a joint attention layer as the embedding projector for the diffusion transformer.

Prefilled Autoregression

When optimizing the autoregressive model, we observe significant error accumulation in the continuous embedding space, attributed to the discrepancy between the training and inference behaviors of autoregressive models. As depicted in Figure 2(a), the model leverages ground-truth preceding



(a) Naive autoregression
Predict next token and reload it as input, token error accumulates

(b) Prefilled autoregression
Prefill input image tokens with special tokens, no error accumulation

Figure 2: (a) The naive autoregressive approach exhibits inconsistent behaviors between training and test phases, leading to error accumulation during inference. (b) We propose a novel strategy that prefills special image tokens during training and testing, which unifies the computational behaviors across both phases and eliminates error accumulation.

tokens for each prediction during teacher-forcing training, whereas at inference time, it relies on autoregressively generated tokens. When processing continuous image embeddings, the model directly predicts biased embeddings and feeds them back as input. This recursive injection of prediction errors propagates and amplifies across subsequent tokens, resulting in suboptimal performance.

Prior research (Li et al. 2024) demonstrates that image token prediction is permutation-invariant. This suggests that each image embedding can be derived solely from the text caption and positional encoding without depending on preceding embeddings. Thus, we propose the prefilled autoregression strategy to mitigate error accumulation, illustrated in Figure 2(b). During training, we initialize all image tokens with N_E learnable embeddings with positional encoding (where N_E is the preset token quantity). During inference, upon predicting the BOI (beginning of image) token, we prefill the input sequence with the learned embeddings. This enforces alignment between training and inference by preventing error-affected predictions from being recycled into the input, thereby eliminating error accumulation.

Dataset Curation

To enable Nexus-Gen with unified visual capabilities, we construct a dataset with 26.3 million samples covering image understanding, generation, and editing tasks. The majority of our dataset derives from publicly accessible open-source repositories (Zhang et al. 2025; Tong et al. 2024a; Zhao et al. 2024a), which is relabeled to improve annotation quality. However, given the image quality limitations (e.g., aesthetic, artifacts) in existing editing datasets, we additionally construct a high-quality image editing dataset, ImagePulse. We will release all training data after data security and legality checks. The complete dataset construction pipeline is provided in the Appendix.

Training Objectives

The training of Nexus-Gen encompasses both its autoregressive model and vision decoder components. The autoregressive model undergoes unified multi-task training, generating outputs comprising both text and image embeddings. The loss function for the text and image embeddings is formulated as follows, where $\lambda_1 = 3$, $\lambda_2 = 1.5$, $\lambda_3 = 1.5$ are hyperparameters that control the loss weights.

$$L = L_{\text{Text}} + L_{\text{Image}} \tag{2}$$

$$= \lambda_1 \cdot L_{\text{CE}} + (\lambda_2 \cdot L_{\text{MSE}} + \lambda_3 \cdot L_{\text{COS}}) \tag{3}$$

For the text tokens, we employ the standard cross-entropy loss for classification, which is defined in Eq. 4. Here, N_T denotes text tokens numbers, $|V|$ represents the vocabulary size, $y_t^{(c)}$ is the ground-truth one-hot encoded token, and $\hat{y}_t^{(c)}$ indicates the predicted probability distribution.

$$L_{\text{CE}} = -\frac{1}{N_T} \sum_{t=1}^{N_T} \sum_{c=1}^{|V|} y_t^{(c)} \log(\hat{y}_t^{(c)}) \tag{4}$$

For the image embeddings, we utilize a composite loss function combining mean squared error and cosine similarity loss (Radford et al. 2021). This combination ensures the preservation of detail fidelity while simultaneously enforcing semantic coherence. The loss functions are defined in Eq. 5 and 6, where N_E is image embeddings numbers, D signifies the embedding dimensionality, $\hat{\mathbf{e}}_i$ and $\mathbf{e}_i$ denote the predicted and ground-truth embeddings, respectively.

$$L_{\text{MSE}} = \frac{1}{N_E \cdot D} \sum_{i=1}^{N_E} \|\mathbf{e}_i - \hat{\mathbf{e}}_i\|_2^2 \tag{5}$$

$$L_{\text{COS}} = -\frac{1}{N_E} \sum_{i=1}^{N_E} \frac{\mathbf{e}_i \cdot \hat{\mathbf{e}}_i}{\|\mathbf{e}_i\|_2 \cdot \|\hat{\mathbf{e}}_i\|_2} \tag{6}$$

The vision decoders for generation and editing tasks are trained separately. They both adopt the MSE loss function of flow matching. Given the diffusion transformer V_θ, target image X_1 in latent space, noise $X_0 \sim N(0, 1)$, conditions C and timestep $t \sim \mu(0, 1)$, the loss function is defined as:

$$L_{\text{Flow}} = \mathbb{E}\left[\|V_\theta(X_t, C, t) - (X_1 - X_0)\|^2\right] \tag{7}$$

Training Strategy

We adopt a multi-stage training strategy for Nexus-Gen. The first stage consists of the unified multi-task pretraining and the aesthetic fine-tuning for autoregressive model, as is shown in Figure 1(b). The second and third stages conduct conditional adaptation for the generation decoder and editing decoder, which is illustrated in Figure 1(c) and (d). All training hyperparameters are listed in the Appendix.

Model	MME-P ↑	MME-C ↑	SEED ↑	MMMU ↑	TextVQA ↑	VQAv2 ↑	RWQA ↑
Qwen2.5-VL-Instruct 7B †	1689.0	640.3	77.4	50.6	77.9	82.3	66.0
EMU2 Chat 34B	-	-	62.8	34.1	<u>66.6</u>	-	-
Seed-X 17B	1457.0	-	66.5	35.6	-	63.4	-
Chameleon 7B	202.7	-	27.2	22.4	-	-	39.0
Chameleon 34B	604.5	-	-	38.8	-	69.6	39.2
EMU3 8B	1243.8	266.1	68.2	31.6	64.7	75.1	57.4
MetaMorph 8B	-	-	71.4	41.8	60.5	-	58.3
Show-O 1.3B	1097.2	-	-	27.4	-	69.4	-
VILA-U 7B	1336.2	-	56.3	32.2	48.3	75.3	46.6
Janus 1.5B	1338.0	-	63.7	30.5	-	-	-
Janus-Pro 1.5B	1444.0	-	68.3	36.3	-	-	-
Janus-Pro 7B	<u>1567.1</u>	-	72.1	41.0	-	-	60.0
LMFusion 16B	1603.7	367.8	72.1	41.7	-	-	<u>60.0</u>
TokenFlow-L 13B	1365.4	257.5	62.6	34.4	54.1	73.9	49.2
TokenFlow-XL 14B	1551.1	<u>371.1</u>	<u>72.6</u>	<u>43.2</u>	62.3	<u>77.6</u>	56.6
Nexus-Gen 7B (Ours)	**1602.3**	**637.5**	**77.1**	**45.7**	**75.5**	**79.3**	**63.7**

Table 1: Evaluation on image understanding benchmarks. We highlight the best results in bold and underline the second-best result. † We evaluate our MLLM base model Qwen2.5-VL-Instruct 7B as a reference baseline.

Multi-Task Pretraining for Autoregressive Model The first stage executes unified pretraining and aesthetic fine-tuning of the autoregressive model. Pretraining utilizes the complete dataset of 26.3 million samples, primarily preserving the autoregressive model's inherent text prediction capacity while establishing a visual embedding prediction functionality. Subsequent aesthetic fine-tuning process employs 4.3 million high-quality samples to optimize Nexus-Gen's visual output quality. This high-quality dataset comprises: (1) premium image generation samples (Zhang et al. 2025; gogoduan 2025; jackyhate 2024; Chen et al. 2025a), (2) our novel ImagePulse dataset augmented with 0.5 million randomly selected instances from other editing corpora (Zhao et al. 2024a; Hui et al. 2024), and (3) 1 million image understanding data (Tong et al. 2024a).

Conditional Adaption for Generation Decoder The second stage fine-tunes the generation decoder to harmonize its conditional inputs with the unified image embedding space via an image reconstruction objective. To preserve visual fidelity, this module is exclusively trained on two million high-quality image generation samples (Zhang et al. 2025; gogoduan 2025; jackyhate 2024; Chen et al. 2025a).

Conditional Adaption for Editing Decoder The third stage fine-tunes the editing decoder using our ImagePulse dataset, as illustrated in Figure 1(d). This decoder processes dual heterogeneous inputs: 324-token fine-grained embeddings extracted from the input image and 81-token semantic embeddings generated by the autoregressive model.

Experiments

Main Results

In this section, we conduct a quantitative evaluation of Nexus-Gen across multiple benchmarks for image understanding, generation, and editing tasks.

Image Understanding For image understanding tasks, we evaluate the model performance on several multimodal understanding benchmarks: MME (Fu et al. 2024), SEED-Bench (Li et al. 2023), MMMU (Yue et al. 2024), TextVQA (Singh et al. 2019), VQAv2 (Goyal et al. 2017), and RealWorldQA (XAI 2024). For baseline comparison, we incorporate the following unified models: EMU2 (Sun et al. 2024b), Seed-X (Ge et al. 2024), Chameleon (Team 2024), Emu3 (Wang et al. 2024), Metamorph (Tong et al. 2024b), Show-O (Tong et al. 2024b), VILA-U (Wu et al. 2024), Janus (Wu et al. 2025), Janus-Pro (Chen et al. 2025b), LMFusion (Shi et al. 2024), and TokenFlow (Qu et al. 2025). As is shown in Table 1, the 7B-parameter Nexus-Gen achieves state-of-the-art performance across all evaluated benchmarks. It outperforms other unified models by significant margins, with particular advantages on MME-Cognition and TextVQA. We further evaluated our MLLM baseline Qwen2.5-VL-Instruct 7B. Empirical results demonstrate that Nexus-Gen endows the base autoregressive model with image generation and editing capabilities without incurring significant loss in image understanding performance, while introducing no additional parameters.

Image Generation For text-to-image generation tasks, we adopt GenEval (Ghosh, Hajishirzi, and Schmidt 2023) as the evaluation benchmark, with assessment metrics covering object fidelity, quantity accuracy, color correctness, attribute matching, and spatial relationships. We compare Nexus-Gen and its instruction-tuned variant Nexus-Gen* (fine-tuned on Blip3o-60k dataset (Chen et al. 2025a)) against unified models including EMU3, SEED-X, Transfusion (Zhou et al. 2025), Show-O (Xie et al. 2024), MetaQuery (Pan et al. 2025), TokenFlow (Qu et al. 2025), Janus (Wu et al. 2025), and Janus-Pro (Chen et al. 2025b). Results in Table 2 reveal that the multi-task jointly-trained Nexus-Gen achieves an overall score of 0.77. After generative-task-specific in-

Method	Single Object ↑	Two Object ↑	Counting ↑	Colors ↑	Position ↑	Color Attribute ↑	Overall ↑
Emu3-Gen	0.99	0.81	0.42	0.80	0.49	0.45	0.66
SEED-X	0.97	0.58	0.26	0.80	0.19	0.14	0.49
Transfusion	-	-	-	-	-	-	0.63
Show-O	0.95	0.52	0.49	0.82	0.11	0.28	0.53
MetaQuery-XL	-	-	-	-	-	-	0.80
TokenFlow-XL	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus	0.97	0.68	0.3	0.84	0.46	0.42	0.61
Janus-Pro 1.5B	0.98	0.82	0.51	0.89	0.65	0.56	0.73
Janus-Pro 7B	**0.99**	0.89	0.59	**0.90**	0.79	**0.66**	0.80
Nexus-Gen (Ours)	0.99	0.86	0.53	0.85	0.78	0.59	0.77
Nexus-Gen* (Ours)	0.97	**0.93**	**0.64**	0.88	**0.83**	0.62	**0.81**

Table 2: Evaluation of image generation ability on GenEval benchmark. We highlight the best results in bold, and underline the second best result. Nexus-Gen underwent joint optimization across all three tasks. We subsequently perform instruction tuning on the Blip3o-60k dataset targeting image generation task, resulting in the optimized Nexus-Gen* variant.

Method	CLIP-T ↑	L1 ↓	CLIP-O ↑	DINO-O ↑
InstructPix2Pix	0.299	0.171	0.832	0.706
MagicBrush	0.309	0.146	0.863	0.750
AnyEdit	0.305	0.141	0.863	0.756
UltraEdit	0.306	0.157	0.841	0.737
OmniGen	0.317	0.154	0.874	0.764
Step1X-Edit	0.317	0.142	0.879	0.779
Nexus-Gen	**0.324**	**0.134**	**0.909**	**0.834**

Table 3: Evaluation of image editing ability on ImagePulse benchmark. We highlight the best results in bold, and underline the second best result.

struction tuning, Nexus-Gen* attains state-of-the-art performance with a score of 0.81. This enhanced model demonstrates significant advantages in Two Object, Counting and Position metrics compared to baseline methods.

Image Editing For image editing evaluation, we utilize the test set with 1,000 randomly sampled cases from ImagePulse dataset (non-overlapping with training data). On this benchmark, we compare Nexus-Gen against state-of-the-art editing models: InstructPix2Pix (Brooks, Holynski, and Efros 2023), MagicBrush (Zhang et al. 2023), AnyEdit (Yu et al. 2025), UltraEdit (Zhao et al. 2024a), OmniGen (Xiao et al. 2025), and Step1X-Edit (Liu et al. 2025). We employ three complementary metric categories: (1) CLIP-T measures the CLIP image-text similarity between edited image and target caption. (2) L1 measures the pixel-level absolute difference between the edited image and groundtruth image. (3) CLIP-O and DINO-O measure the cosine similarity between the edited image and groundtruth image using their CLIP (Radford et al. 2021) and DINO (Caron et al. 2021) embeddings. As evidenced in Table 3, Nexus-Gen demonstrates consistent superiority across all metrics. This performance advantage indicates Nexus-Gen's enhanced capability to faithfully execute editing instructions and generate outputs with significantly improved alignment to both target captions and ground truth images.

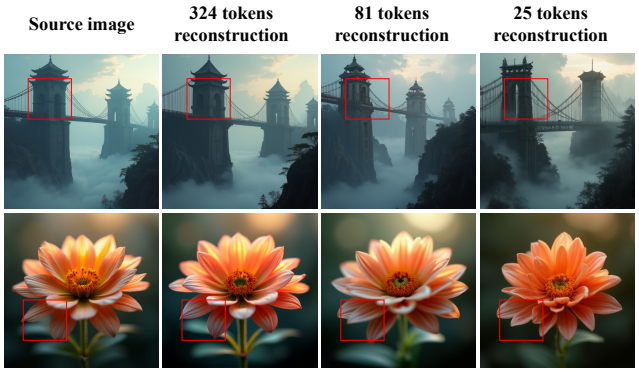


Figure 3: Image reconstruction results of our generation decoder using 81 and 324 image token embedding.

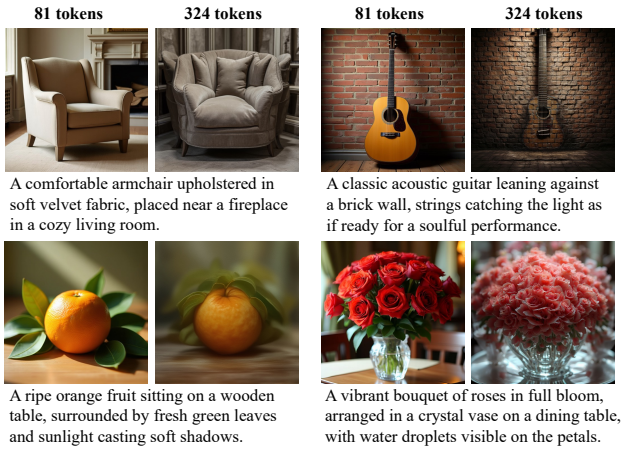


Figure 4: Image generation results from Nexus-Gen trained with 324 and 81 image token embeddings.

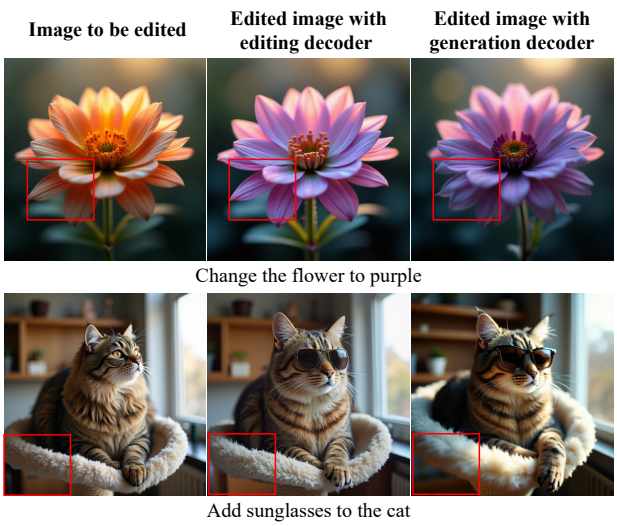


Figure 5: Image editing results of Nexus-Gen with editing and generation decoder.

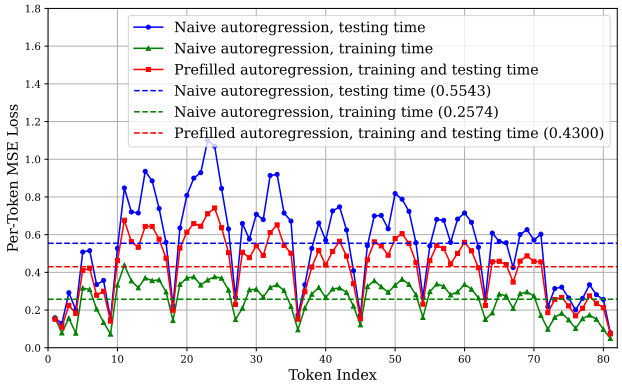


Figure 6: MSE loss comparison between the naive and prefilled autoregression strategy.

Ablation Studies

Trade-offs in Token Quantity The token quantity for image embeddings exhibits a positive correlation with resolution. Higher resolution produces more tokens, allowing embeddings to retain finer visual details. We consider three standard resolutions: 128×128, 256×256 and 512×512, corresponding to 25, 81 and 324 tokens, respectively. We train the generation decoder at these token quantities for image reconstruction, with qualitative results shown in Figure 3. The reconstructions with 81 and 324 tokens successfully preserve global layouts and high-level semantics of source images. Notably, the 324-token approach demonstrates significantly superior detail consistency. In contrast, the 25-token reconstruction exhibits structural distortions and semantic loss relative to the source image.

We further train the autoregressive model at token counts of 81 and 324 and validate image generation quality using above vision decoders. Experimental results in Figure 4 demonstrate that the model trained with 81 tokens effectively generates images aligned with textual semantics. However, the 324-token model exhibits severe semantic repetition in generated images, which exhibited compromised quality. This indicates that autoregressive models struggle to accurately predict such extensive image tokens. Consequently, we select token quantity of 81 as the optimal count for both the autoregressive model and generation decoder.

The Necessity of Editing Decoder Although the generation decoder can be directly applied to image editing tasks, we propose the editing decoder (Figure 1d) to enhance the detail preservation capability in non-edited regions. Figure 5 compares the editing performance of both decoders. Given identical inputs, both solutions successfully execute edit instructions. However, the generation decoder fails to maintain details in non-edited areas due to its 81-token reconstruction constraints. In contrast, the editing decoder synergizes the superior 324-token reconstruction capability with the efficient 81-token embedding prediction, achieving demonstrably higher editing fidelity.

The Impact of Prefilled Autoregression To mitigate the error accumulation issue in naive autoregression paradigm, we propose the prefilled autoregression strategy. Figure 6 compares the MSE losses of image tokens predicted by both approaches. During training, naive autoregression achieves the lowest loss of 0.2574 due to access to preceding ground-truth embedding. During inference however, the autoregressive nature causes progressive error accumulation, elevating the average loss to 0.5543. In contrast, our prefilled autoregression strategy maintains consistent training-inference behavior, achieving a significantly lower inference loss of 0.43.

Conclusion and Future Works

In this work, we present Nexus-Gen, a unified model for image understanding, generation, and editing tasks. The core innovation of Nexus-Gen lies in bridging the language reasoning capabilities of LLMs with the image synthesis power of diffusion models through a unified continuous image embedding space. Furthermore, we identify the error accumulation phenomenon during the autoregressive prediction of continuous embeddings and propose prefilled autoregression strategy to mitigate it. To perform joint optimization across multiple tasks, we curate a large-scale dataset of 26.3 million samples and train Nexus-Gen using a multi-stage strategy, which includes the multi-task pretraining of the autoregressive model and conditional adaptations of the generation and editing decoders. Extensive experiments validate Nexus-Gen's state-of-the-art performance across all tasks. While Nexus-Gen successfully unifies the three image tasks, certain limitations warrant attention. The model exhibits compromised robustness to prompt variations during image generation, and more significantly, its capacity for sophisticated visual reasoning remains unexplored. To address these constraints, we will focus on developing Nexus-Gen's advanced applications in complex tasks such as in-context learning and step-by-step vision-language reasoning.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Chen, J.; Xu, Z.; Pan, X.; Hu, Y.; Qin, C.; Goldstein, T.; Huang, L.; Zhou, T.; Xie, S.; Savarese, S.; et al. 2025a. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025b. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*.

Creative, A. 2024. FLUX-Controlnet-Inpainting. https://github.com/alimama-creative/FLUX-Controlnet-Inpainting.git.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Ge, Y.; Zhao, S.; Zhu, J.; Ge, Y.; Yi, K.; Song, L.; Li, C.; Ding, X.; and Shan, Y. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Ghosh, D.; Hajishirzi, H.; and Schmidt, L. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152.

gogoduan. 2025. flux-laion-aes.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Gu, T.; Yang, K.; Feng, Z.; Wang, X.; Zhang, Y.; Long, D.; Chen, Y.; Cai, W.; and Deng, J. 2025. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. *arXiv preprint arXiv:2504.17432*.

Han, Z.; Mao, C.; Jiang, Z.; Pan, Y.; and Zhang, J. 2024. StyleBooth: Image Style Editing with Multimodal Instruction. *arXiv preprint arXiv:2404.12154*.

Hui, M.; Yang, S.; Zhao, B.; Shi, Y.; Wang, H.; Wang, P.; Zhou, Y.; and Xie, C. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*.

jackyhate. 2024. text-to-image-2M.

Jiang, L.; Yan, Q.; Jia, Y.; Liu, Z.; Kang, H.; and Lu, X. 2025. InfiniteYou: Flexible Photo Recrafting While Preserving Your Identity. In *ICCV*.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Labs, B. F. 2024. FLUX. https://blackforestlabs.ai/announcing-black-forest-labs.

LAION. 2024. laion-high-resolution.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445.

Liu, S.; Han, Y.; Xing, P.; Yin, F.; Wang, R.; Cheng, W.; Liao, J.; Wang, Y.; Fu, H.; Han, C.; et al. 2025. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Ma, Y.; Liu, X.; Chen, X.; Liu, W.; Wu, C.; Wu, Z.; Pan, Z.; Xie, Z.; Zhang, H.; Yu, X.; et al. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7739–7751.

ModelScope. 2025. Diffsynth-Studio.

OpenAI. 2025. Introducing 4o Image Generation.

Pan, J.; Sun, K.; Ge, Y.; Li, H.; Duan, H.; Wu, X.; Zhang, R.; Zhou, A.; Qin, Z.; Wang, Y.; Dai, J.; Qiao, Y.; and Li, H. 2023. JourneyDB: A Benchmark for Generative Image Understanding. arXiv:2307.00716.

Pan, X.; Shukla, S. N.; Singh, A.; Zhao, Z.; Mishra, S. K.; Wang, J.; Xu, Z.; Chen, J.; Li, K.; Juefei-Xu, F.; et al. 2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*.

Qu, L.; Zhang, H.; Liu, Y.; Wang, X.; Jiang, Y.; Gao, Y.; Ye, H.; Du, D. K.; Yuan, Z.; and Wu, X. 2025. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2545–2555.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Shi, W.; Han, X.; Zhou, C.; Liang, W.; Lin, X. V.; Zettlemoyer, L.; and Yu, L. 2024. LlamaFusion: Adapting Pretrained Language Models for Multimodal Generation. *arXiv preprint arXiv:2412.15188*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024a. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; and Wang, X. 2024b. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14398–14409.

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

Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S. C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; et al. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*.

Tong, S.; Fan, D.; Zhu, J.; Xiong, Y.; Chen, X.; Sinha, K.; Rabbat, M.; LeCun, Y.; Xie, S.; and Liu, Z. 2024b. MetaMorph: Multimodal Understanding and Generation via Instruction Tuning. *arXiv preprint arXiv:2412.14164*.

Tuo, Y.; Xiang, W.; He, J.-Y.; Geng, Y.; and Xie, X. 2023. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*.

Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. 2024. Emu3: Next-Token Prediction is All You Need. *arXiv preprint arXiv:2409.18869*.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2022. DiffusionDB: A Large-Scale Prompt Gallery Dataset for Text-to-Image Generative Models. *arXiv:2210.14896 [cs]*.

Wei, C.; Xiong, Z.; Ren, W.; Du, X.; Zhang, G.; and Chen, W. 2024. OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision. *arXiv preprint arXiv:2411.07199*.

Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2025. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12966–12977.

Wu, Y.; Zhang, Z.; Chen, J.; Tang, H.; Li, D.; Fang, Y.; Zhu, L.; Xie, E.; Yin, H.; Yi, L.; et al. 2024. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*.

XAI. 2024. RealWorldQA. https://huggingface.co/datasets/visheratin/realworldqa.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2025. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13294–13304.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Yu, Q.; Chow, W.; Yue, Z.; Pan, K.; Wu, Y.; Wan, X.; Li, J.; Tang, S.; Zhang, H.; and Zhuang, Y. 2025. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26125–26135.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, H.; Duan, Z.; Wang, X.; Chen, Y.; and Zhang, Y. 2025. EliGen: Entity-Level Controlled Image Generation with Regional Attention. *arXiv preprint arXiv:2501.01097*.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36: 31428–31449.

Zhao, H.; Ma, X. S.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; and Chang, B. 2024a. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37: 3058–3093.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; Zhou, W.; and Chen, Y. 2024b. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517.

Zhou, C.; Yu, L.; Babu, A.; Tirumala, K.; Yasunaga, M.; Shamis, L.; Kahn, J.; Ma, X.; Zettlemoyer, L.; and Levy, O. 2025. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*.

Dataset Construction Details

Dataset distribution

We construct a unified dataset covering image understanding, generation and editing tasks with 26.3 million samples. Detailed dataset distribution is presented in Figure 7.

Image Understanding This task is structured with multimodal inputs (image-text pairs) and text-only outputs, which serves as a direct indicator of model's chat and understanding ability. While MLLMs inherently possess such cross-modal reasoning capabilities, this task is still critical during training to prevent capacity degradation. We adopt Cambrian-7M (Tong et al. 2024a) as the data source, a comprehensive dataset spanning multiple domains including optical character recognition, general visual question answering, language, counting, code, math and science tasks. To improve data quality, we re-annotate the answers for all samples with Qwen2.5-VL-72B (Bai et al. 2025).

Image Generation The input for this task is the textual description, and the output is an image. Our data sources comprise Journey DB (Pan et al. 2023), AnyWord (Tuo et al. 2023), Laion-High-Resolution (LAION 2024), EliGen TrainSet (Zhang et al. 2025), FLUX-Aes (gogoduan 2025), FLUX-T2I (jackyhate 2024) and Blip3o-60K (Chen et al. 2025a). To enhance annotation diversity, we employ a dual-captioning paradigm via Qwen2.5-VL-72B, generating both concise captions and elaborate descriptions for each image. During training, we stochastically sample these annotations with stratified ratios (20% concise vs. 80% elaborate) to balance brevity and contextual granularity.

Image Editing The input for editing task consists of an image and its corresponding editing instruction, and the output denotes the edited image. Our data sources encompass datasets such as HQ-Edit (Hui et al. 2024), UltraEdit (Zhao et al. 2024a), OmniEdit (Wei et al. 2024), and StyleBooth (Han et al. 2024). These datasets cover an extensive range of image editing operations, including object-level manipulations, color adjustments, and style transformations. However, these datasets exhibit notable limitations in aesthetic quality. However, they exhibit notable limitations in aesthetic quality and visual fidelity. To address this, we construct the high-quality ImagePulse dataset and integrate it into our training corpus.

Bilingual Annotations Apart from the Chinese samples present in image understanding datasets, all aforementioned datasets are annotated exclusively in English. To endow Nexus-Gen with bilingual (Chinese-English) capabilities for both image generation and editing, additional Chinese image generation and editing samples are incorporated into the dataset. To this end, we perform Chinese re-annotation on several generation subsets (namely EliGen, FLUX-Aes, and FLUX-T2I) as well as the ImagePulse editing dataset. This process yielded a corpus of 2.5 million Chinese training samples, comprising the FLUX-ZH and ImagePulse-ZH subsets illustrated in Figure 7.

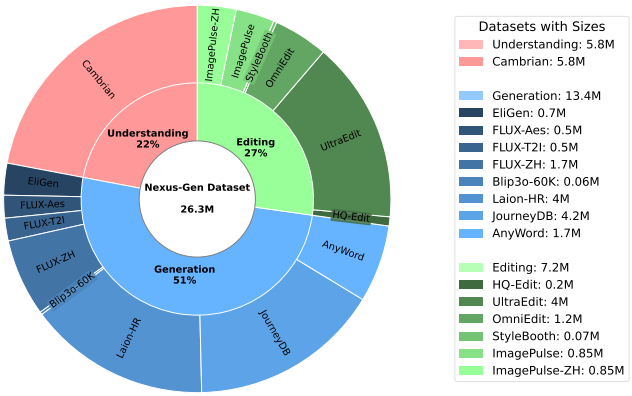


Figure 7: Dataset distribution of our Nexus-Gen dataset.

Construction Pipeline for ImagePulse

Each sample in ImagePulse contains: (1) pristine image pairs synthesized by FLUX.1-Dev (Labs 2024), and (2) semantically-rich editing instruction generated by Qwen2.5-VL-72B (Bai et al. 2025). For diverse editing tasks, the dataset is partitioned into three specialized subsets: Change-Add-Remove, Style Transfer, and FaceID. The workflow for each subset is illustrated in Figure 8

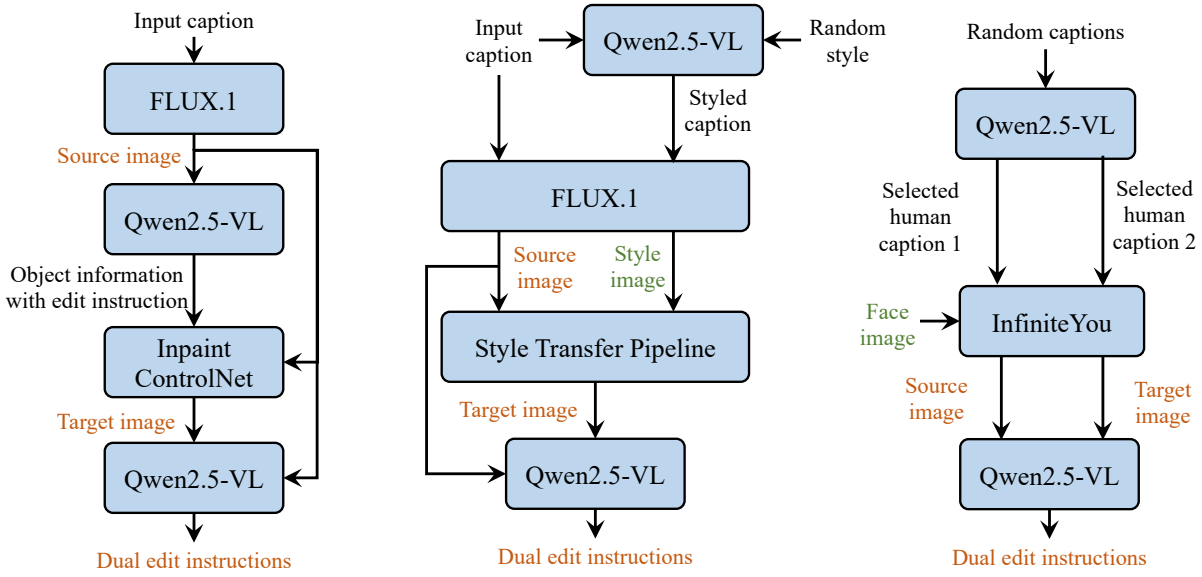
Change-Add-Remove This subset focuses on object-level image manipulations, including modifying object attributes (such as shape, material, and color), adding objects, and removing objects. The dataset construction pipeline is illustrated in Figure 8(a). First, we randomly sample a caption from the DiffusionDB (Wang et al. 2022) dataset and synthesize a source image using the FLUX.1-Dev model. Next, the Qwen2.5-VL model extracts object information, comprising semantic descriptions and spatial locations, from this image and generates corresponding editing instructions. Subsequently, we use the extracted spatial locations and editing instructions to modify specified regions via Inpaint ControlNet (Creative 2024), producing the target image. Finally, to maximize data utility, Qwen2.5-VL generates bidirectional editing instructions between the source and target images.

Style Transfer This subset tackles the problem of image style transfer, with its construction process illustrated in Figure 8(b). First, a randomly selected input caption and a target style prompt are processed by Qwen2.5-VL to generate a style-transformed caption. Subsequently, the original input caption and the generated style-transformed caption are leveraged to synthesize the source image and the style image, respectively. Crucially, the style image exhibits significant structural divergence from the source image, rendering it unsuitable as the target image. To derive the target image, we devise a Style Transfer Pipeline integrating ControlNet, SDXL, InstantStyle, and IP-Adapter models. This pipeline effectively fuses the structural framework of the source image with the stylistic properties of the style image. Finally, we generate corresponding dual editing instructions.

FaceID The unified architecture of Nexus-Gen facilitates tackling particularly challenging conceptual editing tasks by

Training Phase	Multi-task Pretraining for Autoregressive Model		Conditional Adaption for Vision Decoder	
Training Target	Large Scale Pretraining	Aesthetic Fine-tuning	Generation Decoder	Editing Decoder
Learning Rate	1e-5	1e-5	1e-5	1e-5
LR Scheduler	Cosine	Cosine	Constant	Constant
Batch Size	512	512	128	128
Total Steps	100 K	10 K	20 K	10 K
Warm-up Steps	7500	800	100	100
Total Samples (Million)	26	4	2	1
Data Ratio (Und:Gen:Edit)	1:2:1	1:2:1	0:1:0	0:0:1

Table 4: Detailed hyperparameters for training Nexus-Gen. Data ratio refers to the ratio of image understanding data, generation data and editing data.



(a) Change-Add-Remove subset (b) Style Transfer subset (c) FaceID subset

Figure 8: The dataset construction pipeline for three subsets of ImagePulse.

harnessing advanced image generation capabilities. To validate this capability, we construct the FaceID subset, which focuses on executing high-variation conceptual edits while preserving subject identity. The dataset construction process is illustrated in Figure 8(c). Initially, Qwen2.5-VL selects two captions containing human descriptions from a pool of randomly sampled image captions. Subsequently, a facial image is randomly sampled from the CelebA-HQ-Face (Liu et al. 2015) dataset. Using these two captions and the facial image, we synthesize the source and target images via the InfiniteYou (Jiang et al. 2025) model. These images exhibit identical subject identity while differing significantly in pose and scene context. Finally, dual editing instructions corresponding to the source and target images are generated by Qwen2.5-VL.

Implementation Details

Nexus-Gen employs Qwen2.5-VL as its autoregressive model and FLUX.1-Dev as the generation and editing decoder, utilizing a multi-stage training strategy to optimize each component separately. The autoregressive model undergoes training within the ms-swift (Zhao et al. 2024b) framework, comprising pretraining followed by an aesthetic fine-tuning stage. The generation and editing decoders are trained using Diffsynth-Studio (ModelScope 2025). Detailed training hyperparameters are provided in Table 4. During inference, the generation decoder utilizes classifier-free guidance with a scale of 3.0.

More Qualitative Results

In this section, we present additional qualitative results for Nexus-Gen, focusing on image generation and editing tasks.

Image Generation

Figure 9 showcases representative high-fidelity images synthesized by Nexus-Gen, demonstrating the model's capability to accurately interpret semantic information from textual descriptions and translate them into visually coherent outputs. Owing to the incorporation of bilingual image generation datasets, Nexus-Gen is capable of processing inputs and generating outputs in both English and Chinese.

Image Editing

The image editing capabilities of Nexus-Gen are demonstrated in Figure 10, which exhibits seamless handling of diverse workflows including subject addition, removal, replacement, color alteration, and style transfer. It can be observed that Nexus-Gen demonstrates remarkable proficiency in both preserving non-edited regions and executing editing instructions.

Limitations

Despite its capabilities in image understanding, generation, and editing, Nexus-Gen exhibits distinct limitations. First, with a total training dataset of 26 million samples, its scale remains substantially smaller than either specialized single-task models or unified counterparts trained on hyper-scale datasets. Consequently, the model may exhibit sensitivity to image generation prompts and often requires specific instruction templates for optimal editing performance. Second, unlike VAE latent spaces that enable precise pixel-level reconstruction, Nexus-Gen's unified image space operates primarily at the semantic feature level, resulting in inherent reconstruction fidelity limitations. Third, the visual reasoning capabilities of Nexus-Gen remains unexplored.

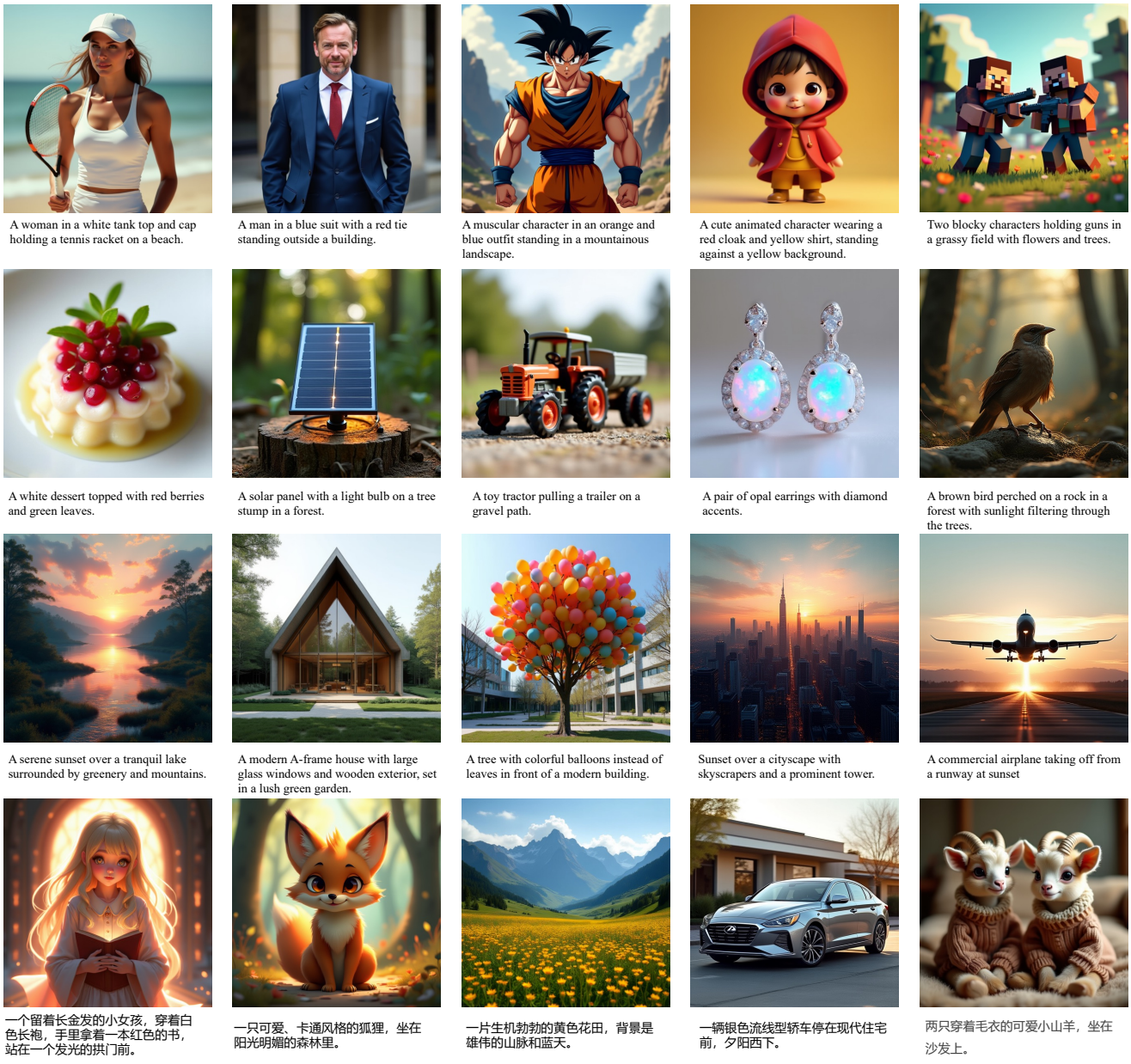


Figure 9: Qualitative image generation results of Nexus-Gen.

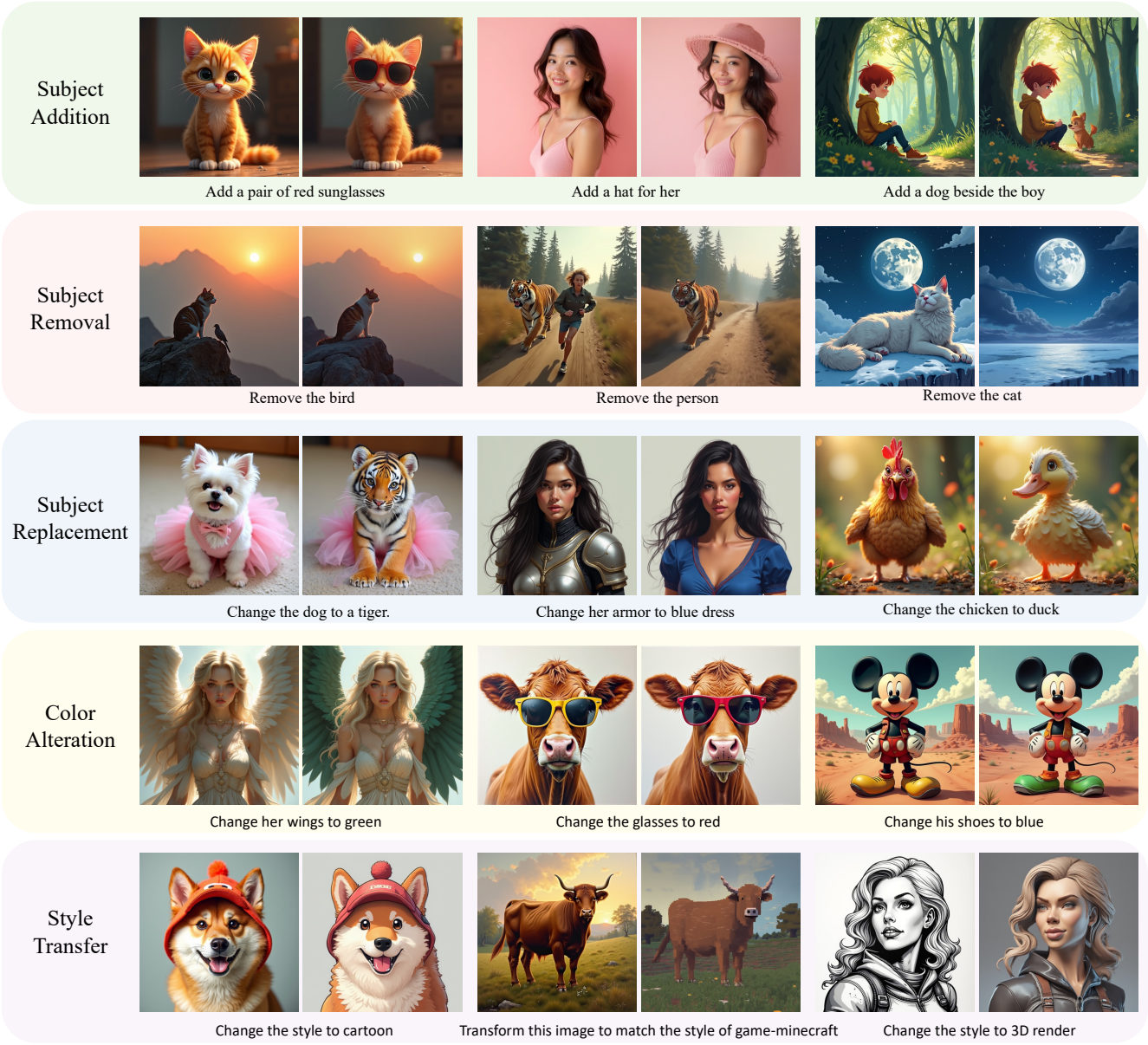


Figure 10: Qualitative image editing results of Nexus-Gen.