EchoGen: Generating Visual Echoes in Any Scene via Feed-Forward Subject-Driven Auto-Regressive Model

Anonymous authors
Paper under double-blind review

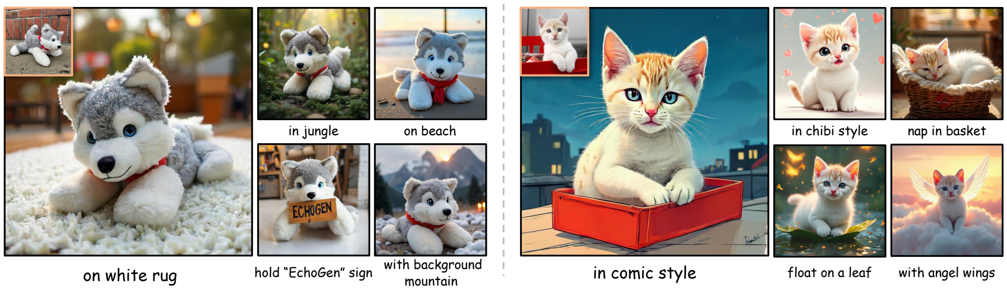


Figure 1: **Feed-forward subject-driven generation by EchoGen**. By employing a visual autoregressive paradigm, EchoGen achieves both high-quality image synthesis with lower latency, preserving intricate subject identity with exceptional efficiency.

Abstract

Subject-driven generation is a critical task in creative AI; yet current state-of-the-art methods present a stark trade-off. They either rely on computationally expensive, per-subject fine-tuning, sacrificing efficiency and zero-shot capability, or employ feed-forward architectures built on diffusion models, which are inherently plagued by slow inference speeds. Visual Auto-Regressive (VAR) models are renowned for their rapid sampling speeds and strong generative quality, making them an ideal yet underexplored foundation for resolving this tension. To bridge this gap, we introduce **EchoGen**, a pioneering framework that empowers VAR models with subject-driven generation capabilities. The core design of EchoGen is an effective dual-path injection strategy that disentangles a subject's high-level semantic identity from its low-level fine-grained details, enabling enhanced controllability and fidelity. We employ a semantic encoder to extract the subject's abstract identity, which is injected through decoupled cross-attention to guide the overall composition. Concurrently, a content encoder captures intricate visual details, which are integrated via a multi-modal attention mechanism to ensure high-fidelity texture and structural preservation. To the best of our knowledge, EchoGen is the first feed-forward subject-driven framework built upon VAR models. Both quantitative and qualitative results substantiate our design, demonstrating that EchoGen achieves subject fidelity and image quality comparable to state-of-the-art diffusion-based methods with significantly lower sampling latency.

1 Introduction

The rapid evolution of text-to-image synthesis models (Saharia et al., 2022; Rombach et al., 2022; Batifol et al., 2025; Esser et al., 2024) has catalyzed a variety of novel applications (Zhang et al., 2023), among which subject-driven generation stands out as an important task. This task aims to accurately depict a specified subject within diverse, user-defined scenes described through text prompts, while rigorously upholding the subject's core identity. The early approaches (Ruiz et al., 2023; Gal et al., 2022; Kumari et al., 2023) introduced a test-time fine-tuning paradigm that optimizes a large pretrained model using a few images for each new subject. Although effective in preserving identity to some extent, this per-subject optimization process is computationally expensive, demanding at least hundreds of training iterations and substantial GPU resources, ultimately resulting in a distinct model checkpoint for each subject. These limitations significantly hinder the practicality and scalability of the test-time fine-tuning paradigm in real-world applications.

To improve efficiency and practicality, a new class of feed-forward approaches has recently emerged (Li et al., 2023; Pan et al., 2024; Ye et al., 2023; Tan et al., 2025; Shin et al., 2025) based on diffusion models (Rombach et al., 2022; Podell et al., 2024; Batifol et al., 2025). Instead of fine-tuning on a small set of images for each new subject, feed-forward approaches perform a single, large-scale supervised fine-tuning on a vast dataset composed of triplets (text, reference image, target image). The model is trained to learn a generalizable mapping from a subject image to the snapshot version in the specified scene. The single process of pretraining enables zero-shot generation at inference time–a novel subject can be synthesized immediately without any subject-specific fine-tuning, significantly reducing the initial setup cost and decreasing generation latency by eliminating the need for test-time optimization. Nevertheless, these methods still inherit the computational demands of the underlying diffusion models due to the iterative denoising process.

Inspired by autoregressive generation in language models (Radford et al., 2018; Achiam et al., 2023), autoregressive visual generation (Esser et al., 2021; Ramesh et al., 2021; Sun et al., 2024) has emerged as a compelling alternative to diffusion models. Unlike diffusion's iterative denoising, autoregressive models synthesize content sequentially, token by token. This paradigm is further advanced by the Visual Autoregressive (VAR) model (Tian et al., 2024; Han et al., 2025), which employs a coarse-to-fine *next-scale* generation strategy instead of traditional *next-token* generation. It first generates tokens for the global composition and then renders fine-grained details, capturing a complete hierarchical representation from structure to texture. The novel paradigm allows VAR to achieve superior performance compared to traditional autoregressive models, outperforming top-tier diffusion models while offering faster inference speed. Despite the inherent suitability of the autoregressive paradigm for fine-grained conditioning, its potential for controllable generation, especially in the feed-forward, subject-driven context, remains largely untapped compared to the wealth of research on diffusion-based methods. This critical gap severely limits the practical applicability of VAR models, hindering their adoption in real-world scenarios where subject control is paramount.

In this work, we aim to bridge this gap by leveraging the inherent advantages of VAR to build an effective, scalable, and highly controllable system for subject-driven image synthesis. We propose *EchoGen*, the first efficient **feed-forward** autoregressive framework that generates faithful visual renditions of a given subject in arbitrary scenes. At the core of EchoGen is a *dual-path* injection mechanism that disentangles semantic features from fine-grained details. We inject high-level fine-grained semantic features extracted by a semantic encoder based on the pretrained vision foundation model (DINOv2 (Oquab et al., 2024)) into the decoupled cross-attention layers (Kumari et al., 2023) to bring structural and stylistic coherence while avoiding drift in prompt following. To enable global semantic conditioning, we prepend the global semantic embedding extracted from DINOv2 as a prefix and subsequently infuse it via Adaptive LayerNorm, thereby steering the overall semantic generation. However, generating with semantic features alone often misses low-level details. To complement these features, a second pathway employs a pretrained content encoder (FLUX.1-dev VAE (Batifol et al., 2025)) to extract fine-grained image features, which are incorporated via a multi-modal attention module, ensuring faithful reconstruction of local textures and details. To preserve the generative capabilities of the pretrained VAR model, we adopt a parameter-efficient fine-tuning strategy that freezes the backbone and only updates key components within the subject injection modules. Extensive quantitative and qualitative evaluations on DreamBench (Ruiz et al., 2023) benchmark and human evaluation demonstrate that EchoGen achieves subject fidelity, text alignment, and image quality comparable to and even exceeding state-of-the-art diffusion-based methods, while exhibiting lower sampling latency.

Our principal contributions can be summarized as follows:

- We introduce EchoGen, the first feed-forward, efficient, subject-driven generation framework built upon a visual autoregressive model. This establishes a compelling new paradigm for controllable subject-driven synthesis beyond the dominant diffusion-based approaches.
- We propose a novel dual-path injection strategy that disentangles the identity of a subject into high-level semantics and fine-grained details. By injecting these features through separate pathways within a parameter-efficiently tuned model, EchoGen achieves faithful subject representation across diverse scenes.

- Extensive experiments demonstrate that EchoGen achieves subject fidelity, text alignment, and image quality that are competitive with or superior to state-of-the-art diffusion-based methods with much faster inference speed.

2 Related Works

2.1 Autoregressive Image Generation

Unlike diffusion-based methods that synthesize images via iterative denoising, the autoregressive paradigm models image distributions by sequentially predicting visual tokens conditioned on the preceding context. This approach evolves from inefficient and low quality early pixel-level methods (Van den Oord et al., 2016; Salimans et al., 2017) to a dominant two-stage framework that first compresses images into discrete tokens and then models their distribution utilizing Transformer (Esser et al., 2021). This paradigm substantially improves generation fidelity and efficiency, underpinning advances in text-to-image synthesis (Ramesh et al., 2021; Yu et al., 2022b) and controllable generation (Li et al., 2025). Subsequent work further refines it by improving image tokenizers (Yu et al., 2022a; Mentzer et al., 2024), exploring continuous representations with diffusion modeling (Li et al., 2024a; Fan et al., 2025), or adapting large language models for visual generation (Sun et al., 2024; Wu et al., 2024). To mitigate structural degradation induced by the fixed raster-scan order, Visual Autoregressive (VAR) models (Tian et al., 2024) introduce a hierarchical coarse-to-fine strategy that progressively refines fine-grained details by next-scale prediction. The following version Infinity (Han et al., 2025) extends the VAR model to text-to-image generation, achieving superior quality with significantly lower sampling latency than diffusion models. While existing works extend VAR to controllable generation (Yao et al., 2024; Li et al., 2024b; Chung et al., 2025), feed-forward subject-driven personalization remains underexplored, limiting the practical applicability of the VAR framework.

2.2 Subject-driven Image Generation

Test-time fine-tuning methods. Diffusion models (Ho et al., 2020; Rombach et al., 2022) have achieved remarkable success in high-fidelity text-to-image (T2I) synthesis (Podell et al., 2024; Esser et al., 2024; Batifol et al., 2025). For subject-driven tasks, relying solely on text prompts is often insufficient to preserve the defining characteristics of specific subjects. To address this, pioneering methods (Gal et al., 2022; Ruiz et al., 2023; Kumari et al., 2023) introduce customization by fine-tuning on a small set of reference images for each target subject. While these approaches can capture intricate details and deliver high fidelity to some extent, their dependence on per-subject optimization remains time-consuming and computationally demanding, which limits practical use.

Feed-forward subject-driven approaches. To overcome the efficiency limitations of per-subject optimization, feed-forward methods have been developed (Wei et al., 2023; Zeng et al., 2024; Patel et al., 2024; Ma et al., 2024; Wang et al., 2025). These models are trained once to condition on subject features from vision encoders, enabling fast, zero-shot synthesis for novel subjects. Early works such as BLIP-Diffusion (Li et al., 2023) jointly fine-tune the denoising network with multimodal alignment modules while suffering from inadequate fidelity and image quality. To mitigate the high computational cost of full model tuning, parameter-efficient strategies (Pan et al., 2024; Ye et al., 2023; Tan et al., 2025; Zhang et al., 2025; Wu et al., 2025) incorporate lightweight modules such as LoRA (Hu et al., 2022) or adapters. These modules inject reference features into the diffusion transformer, typically via attention mechanisms, while keeping most pretrained weights frozen. However, since these methods all rely on diffusion backbones, they inherit the substantial inference latency of the iterative denoising process, which constrains their practical deployment.

3 Preliminary of visual autoregressive modeling

Autoregressive models (Esser et al., 2021) reframe image synthesis as a sequential token prediction, under the *next-token* prediction. The image is first tokenized into a discrete feature map using a visual tokenizer $\mathcal{E}$ and then flattened into a one-dimensional sequence, typically following the raster scan order. The model is then trained to predict each token x_i given the preceding tokens $(x_1, ..., x_{i-1})$ and the condition c, factorizing the sequence distribution as $p(x_1, \ldots, x_N|c) = \prod_{i=1}^{N} p(x_i|x_1, \ldots, x_{i-1}, c)$. However, the vanilla next-token paradigm with fixed raster order induces structural degradation and insufficient modeling.

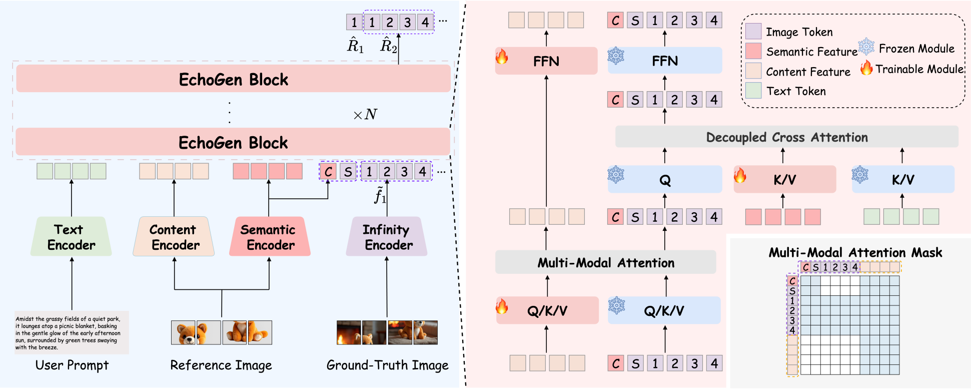


Figure 2: **Overview of the EchoGen architecture**. The left panel illustrates the overall model framework with dual-path subject injection, while the right panel provides a detailed schematic of the EchoGen block with a carefully designed attention mask applied in the Multi-Modal Attention module to avoid feature leakage. C denotes the global semantic token extracted from the semantic encoder, which is prepended to the input sequence. S represents the start token for the first-scale generation. Adaptive Layer Normalization modules in the EchoGen blocks are omitted for clarity.

Visual autoregressive modeling (Tian et al., 2024) addresses the above issues by shifting the prediction paradigm from *next-token* to *next-scale*: instead of predicting one token at a time, it predicts entire token maps at *progressively increasing resolutions*. The visual encoder $\mathcal{E}$ first maps an image I to latent F, and then produces K multi-scale token maps $(r_1, ...r_K)$ with increasing resolutions $h_k \times w_k$ by applying a residual vector quantizer. A GPT-style Transformer begins from the generation of the 1×1 map r_1 and autoregressively predicts each subsequent scale given prior scales and condition c, achieving generation from global structure to fine details, which is formulated as:

$$p(r_1, \ldots, r_K | c) = \prod_{k=1}^{K} p(r_k | r_1, \ldots, r_{k-1}, c). \tag{1}$$

This scale-wise coarse-to-fine paradigm is well suited for scalable text-to-image generation. The text-to-image generation model Infinity (Han et al., 2025) leverages bitwise quantization to expand the vocabulary size under the next-scale paradigm, reporting state-of-the-art performance with reduced sampling latency compared to diffusion baselines. In this paper, to bypass the cumbersome per-subject fine-tuning and the heavy computational cost during inference, we propose a novel feed-forward framework based on VAR models, featuring a single parameter-efficient fine-tuning phase.

4 EchoGen

4.1 Overall Framework

We are seeking a novel feed-forward framework for subject-driven generation built upon Infinity, based on the proposed **EchoGen** block with effective dual-path subject information injection, in which a content encoder and a semantic encoder cooperate to provide comprehensive subject features from both sides of a coin. The overview of the EchoGen architecture and its basic block is illustrated in Figure 2. Before subject injection, to ensure robustness against background noise that may interfere with subject injection, a pipeline based on the multi-modality model Qwen2.5-VL (Bai et al., 2025) and the open segmentation model GroundingDINO (Liu et al., 2024) is carefully designed to segment the subject from complex scenes. Given the segmented subject image, our EchoGen model is trained using a parameter-efficient methodology that freezes the pretrained backbone while fine-tuning only newly introduced attention modules. During inference, we apply flexible subject-text classifier-free guidance for explicit control over the trade-off between subject fidelity and textual alignment, enabling versatile and controllable generation.

4.2 Dual-Path Subject Injection

Semantic feature injection for identity preservation. The semantic feature, which captures abstract characteristics, provides a representation that is critical for avoiding the identity drift common in subject-driven generative models. Following this principle, we introduce a bifurcated injection strategy that targets both the fine-grained and global levels of the generative process. For fine-grained conditioning, we employ the pretrained DINOv2 vision encoder to extract patch-level semantic embeddings. These embeddings are synergistically integrated with the original textual conditioning via a decoupled cross-attention mechanism (Kumari et al., 2023). Our decoupled cross-attention mechanism operates on query features $\mathcal{Z}$, conditioning them on both the text embedding c_t and the fine-grained semantic features c_s, formulated as follows:

$$\begin{aligned} \mathcal{Q} = \mathcal{Z}\, W^{q},\, \mathcal{K} = \text{concat}\left(c_s\, W_s^k,\, c_t\, W_t^k\right),\, \mathcal{V} = \text{concat}\left(c_s\, W_s^v,\, c_t\, W_t^v\right), \\ \mathcal{Z}' = \text{Attention}\left(\mathcal{Q}, \mathcal{K}, \mathcal{V}\right) = \text{Softmax}\left(\mathcal{Q}\mathcal{K}^{\top}/\sqrt{d}\right)\mathcal{V}, \end{aligned} \tag{2}$$

where W^q is the query projector, (W_t^k, W_t^v) and (W_s^k, W_s^v) are two distinct sets of (k, v) projectors to embed text prompting c_t and semantic injection c_s, respectively. The resulting key and value pairs for each condition are concatenated to form the final context vectors $\mathcal{K}$ and $\mathcal{V}$. We keep the projectors for text prompting (W_t^k, W_t^v) and the query projector W^q frozen while exclusively optimizing the key and value projectors (W_s^k, W_s^v) that map the semantic features of the reference images, enabling an alignment mapping from the semantic visual space to the generator's latent space without perturbing the pretrained knowledge.

Moreover, we prepend the DINOv2 global semantic token C to the input sequence to impose holistic semantic guidance. At the same time, this global token also serves as a condition for the Adaptive Layer Normalization (AdaLN) layer in the proposed EchoGen block, following (Han et al., 2025). The infusion of fine-grained and global semantics ensures comprehensive semantic-informed generation, promoting fine-grained fidelity and global structural coherence.

Content feature injection for detail preservation. While the semantic embeddings provide a robust identity preservation, their high abstraction leads to generation with insufficient subject details. To achieve high fidelity of the subject's content, we complement it with a content feature infusion mechanism. To be specific, EchoGen employs the FLUX.1-dev VAE to extract low-level content features c_c, which are then integrated via the multi-modal attention. The generation process is then steered by a carefully designed attention operation: generated tokens have unobstructed access to the reference tokens, allowing them to distill fine-grained visual cues on demand; conversely, a causal mask renders the reference tokens oblivious to the generated sequence, which is a critical constraint for ensuring the autoregressive sampling trajectory. This masking schema is precisely demonstrated in the lower-right inset of Figure 2. Specifically, given the generated token sequence $\mathcal{Z}$ and the detailed content condition c_c, the multi-modal attention utilizes separate linear projections (W^q, W^k, W^v) for $\mathcal{Z}$ and (W_c^q, W_c^k, W_c^v) for the condition c_c, with the applied attention mask Mask, and then calculate the generated sequence $\mathcal{Z}'$ and condition ${c_c}'$ via:

$$\begin{aligned} \mathcal{Q} = \text{concat}\left(\mathcal{Z}\, W^{q},\, c_c\, W_c^q\right),\, \mathcal{K} = \text{concat}\left(\mathcal{Z}\, W^{k},\, c_c\, W_c^k\right),\, \mathcal{V} = \text{concat}\left(\mathcal{Z}\, W^{v},\, c_c\, W_c^v\right), \\ \mathcal{Z}', c_c' = \text{Attention}\left(\mathcal{Q}, \mathcal{K}, \mathcal{V}, \text{Mask}\right) = \text{Softmax}\left(\text{Mask}\left(\mathcal{Q}\mathcal{K}^{\top}/\sqrt{d}\right)\right)\mathcal{V}. \end{aligned} \tag{3}$$

The pathways for the generated token sequence remain frozen, while exclusively parallel attention projectors (W_c^q, W_c^k, W_c^v) and FFN modules for processing content features are optimized.

Through this dual-path subject injection strategy, our model faithfully preserves the salient visual characteristics of the reference image while simultaneously maintaining a strong adherence to the provided text instructions.

4.3 Subject Segmentation

A common challenge in real-world scenarios is that user-provided reference images comprised of the subject of interest within visually complex backgrounds may harm the performance of subject injection. To mitigate this issue, we employ a subject segmentation pre-processing pipeline, illustrated in Figure 3. First, the Qwen2.5-VL (Bai et al., 2025) vision-language model identifies the subject's semantic identity, producing a descriptive text prompt. This prompt

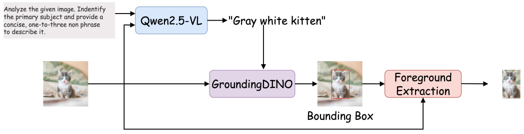

Analyze the given image. Indentify the primary subject and provide a concise, one-to-three non phrase to describe it.
Qwen2.5-VL
"Gray white kitten"
GroundingDINO
Bounding Box
Foreground Extraction

Figure 3: **The pipeline of subject segmentation**.

is then used to condition the GroundingDINO (Liu et al., 2024) model for precise subject localization and bounding box generation. The foreground region is subsequently cropped according to this bounding box, while the surrounding unrelated regions are explicitly discarded and replaced with a uniform white background. This process ensures that subsequent feature injection operates attentively on the isolated representation of the referred subject.

4.4 Sampling with Subject-Text Classifier-free Guidance

Classifier-Free Guidance (CFG) (Ho & Salimans, 2021) has become a cornerstone technique for enhancing conditional control in generative models, especially in diffusion models. Its core principle is to amplify the conditional signal by extrapolating from an unconditional prediction towards a conditional one, thereby improving condition following at the cost of some diversity. Recently, many autoregressive models (Chang et al., 2023; Tian et al., 2024) have also incorporated CFG into their frameworks. In this work, we further enhance the influence of the text embedding c_t and the subject condition c_s, c_c within the CFG scheme for subject-driven generation. During training, we independently replace the text condition c_t with an unconditional token $\varnothing_t$ and the image condition c_s, c_c with unconditional embeddings $\varnothing_s, \varnothing_c$, each with a probability of 10%. During inference, assuming the independence between the text condition c_t and the image condition c_s, c_c, we compute the final logits predicted by EchoGen via a flexible guidance rule that integrates both controls:

$$\hat{l} = l(\varnothing_t, \varnothing_s, \varnothing_c) + \gamma_t \times (l(c_t, \varnothing_s, \varnothing_c) - l(\varnothing_t, \varnothing_s, \varnothing_c)) + \gamma_I \times (l(c_t, c_s, c_c) - l(c_t, \varnothing_s, \varnothing_c)), \quad (4)$$

where l denotes the Transformer output logits, and γ_t together with γ_I are hyperparameters that govern the guidance scales. This dynamic text-subject guidance not only strengthens the influence of text embeddings and image prompts, thereby improving generation performance, but also provides a flexible mechanism to balance text alignment with the reference preservation.

5 Experiment

5.1 Setup

Datasets. We conduct experiments on the Subjects200K dataset (Tan et al., 2025), a substantial collection comprising about 256,000 triplets of text prompts, reference images, and target images. The corpora were generated using GPT-4o and FLUX.1-dev, and with image resolutions larger than 500×500. For EchoGen-0.1B training, both the reference and target images are resized and center-cropped to 256×256. To enable high-resolution generation for EchoGen-2B training, we avoid direct interpolation, which may introduce undesirable artifacts; instead, we upscale the images to 1024×1024 using the PiSA-SR super-resolution model (Sun et al., 2025).

Training details. Our training protocol largely follows Infinity (Han et al., 2025). We train EchoGen for 400k iterations, utilizing the AdamW (Loshchilov & Hutter, 2017) optimizer with a global batch size of 128, setting the base learning rate as 3×10^{-5} and the momentum parameters $(\beta_1, \beta_2) = (0.9, 0.97)$. To stabilize fine-tuning, we apply a reduced learning rate of 3×10^{-6} to the multi-modal attention parameters. More training details can be found in the appendix 7.3.1.

Evaluation. Following prior works (Ruiz et al., 2023; Li et al., 2023), we evaluate our approach in terms of subject fidelity and text alignment on the DreamBench benchmark (Ruiz et al., 2023). Subject fidelity is measured by the cosine similarity between the generated and reference images using both CLIP (Radford et al., 2021) image embeddings (CLIP-I) and DINO (Zhang et al., 2022) features (DINO). Text alignment is assessed via the CLIP cosine similarity between the generated image and its corresponding input prompt (CLIP-T). DreamBench, comprising real-world images with prompt annotations, includes 30 unique subjects, each paired with 25 distinct prompts. Following the standard protocol, we generate four images for each prompt-subject pair, yielding a total of 3,000 images for evaluation.

5.2 Main Results

We compare EchoGen with three categories of prior works: (1) test-time fine-tuning methods that require per-subject optimization; (2) unified generation models with large-scale pre-training; and (3) feed-forward approaches that share the same paradigm as ours and constitute our most baselines.

Quantitative results. We benchmark EchoGen performance against contemporary subject-driven diffusion-based methods in the DreamBench dataset (Ruiz et al., 2023), with quantitative results

Method	Base Model	DINO↑	CLIP-I↑	CLIP-T↑	Latency ↓
		Test-time Fine-tuning			
Textual-Inversion (Gal et al., 2022)	SD-v1.5	0.569	0.780	0.255	50min
DreamBooth (Ruiz et al., 2023)	SD-v1.5	0.668	0.803	0.305	15min
BLIP-Diffusion (Li et al., 2023)	SD-v1.5	0.670	0.805	0.302	-
AR-Booth (Chung et al., 2025)	Infinity-2B	**0.750**	**0.808**	0.269	2.8h
		Unified Generation			
OmniGen (Xiao et al., 2025)	OmniGen	**0.693**	0.801	0.315	93.4s
		Feed-Forward			
ELITE (Wei et al., 2023)	SD-v1.4	0.621	0.771	0.293	**11.0s**
Re-Imagen (Chen et al., 2023)	Imagen	0.600	0.740	0.270	-
BLIP-Diffusion (Li et al., 2023)	SD-v1.5	0.594	0.779	0.300	-
λ-Eclipse (Patel et al., 2024)	Kan-v2.2	0.613	0.783	0.307	-
BootPIG (Purushwalkam et al., 2024)	SD-v2.1	0.674	0.797	0.311	-
MS-Diffusion (Wang et al., 2025)	SDXL	0.671	0.792	0.321	39.6s
IP-Adapter (Ye et al., 2023)	SDXL	0.613	**0.810**	0.292	16.9s
IP-Adapter (Ye et al., 2023)	FLUX.1-dev	0.561	0.725	**0.351**	-
OminiControl (Tan et al., 2025)	FLUX.1-dev	0.684	0.799	0.312	27.5s
EasyControl (Zhang et al., 2025)	FLUX.1-dev	0.652	0.789	**0.325**	47.6s
EchoGen-0.1B	Infinity-0.1B	0.675	0.806	0.321	**0.5s**
EchoGen-2B	Infinity-2B	**0.755**	**0.837**	**0.324**	**5.2s**

Table 1: **Quantitative comparisons on DreamBench (Ruiz et al., 2023)**. We highlight the best, second-best, and third-best values for each metric. The results indicate that EchoGen attains performance on par with diffusion-based approaches while delivering substantially faster sampling.

Method	Subject Fidelity↑	Text Alignment↑	Photorealism↑
OmniGen (Xiao et al., 2025)	0.15	0.13	0.09
IP-adapter (Ye et al., 2023)	0.21	0.05	0.14
OminiControl (Tan et al., 2025)	0.12	0.21	0.15
EasyControl (Zhang et al., 2025)	0.15	**0.31**	0.28
EchoGen-2B	**0.37**	0.30	**0.34**

Table 2: **Human evaluation**. We compare our method with previous approaches based on three aspects: text alignment, subject fidelity, and photorealism.

summarized in Table 1. EchoGen achieves performance that superior to leading diffusion-based approaches in the core metrics of subject fidelity and text alignment, and demonstrates balanced performance across evaluation axes. In contrast, several baselines, such as IP-Adapter (Ye et al., 2023) exhibit significant weaknesses in specific metrics. Furthermore, the adoption of the visual autoregressive paradigm provides a clear efficiency advantage: EchoGen's inference latency for a 1024×1024 image is under 6 seconds, representing a significant acceleration over the more than 10 seconds required by its diffusion-based counterparts. Overall, these results indicate that EchoGen combines strong generative quality with markedly improved efficiency, offering a competitive alternative for subject-driven synthesis.

Qualitative results. Figure 4 presents a rigorous qualitative comparison with prominent diffusion-based frameworks, revealing substantial advantages of our model in both subject fidelity and prompt correspondence. EchoGen exhibits the ability to render high-fidelity details, such as the precise reconstruction of the teapot spout and the nuanced texture of the sloth plushie, and we attribute this capability to our dual-path semantic-content feature injection design. In contrast, baselines including IP-Adapter (Ye et al., 2023) and OminiControl (Tan et al., 2025) exhibit characteristic failure cases, corroborating EchoGen's robustness. EchoGen also demonstrates more consistent compliance with

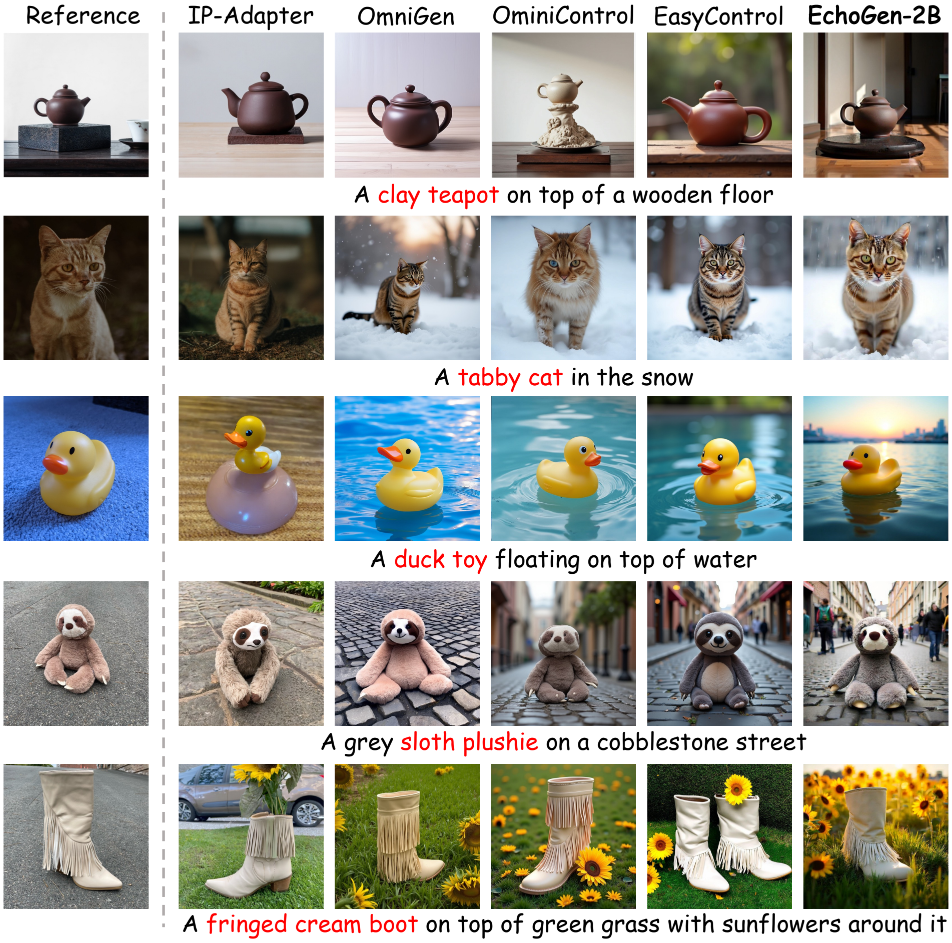

Reference
IP-Adapter
OmniGen
OminiControl
EasyControl
EchoGen-2B
A clay teapot on top of a wooden floor
A tabby cat in the snow
A duck toy floating on top of water
A grey sloth plushie on a cobblestone street
A fringed cream boot on top of green grass with sunflowers around it

Figure 4: **Qualitative comparison with diffusion-based methods on DreamBench (Ruiz et al., 2023)**. For a fair comparison, we adopt the default sampling settings for all baseline models.

textual prompts, avoiding the language deviations observed in the generations of the duck toy and cat instances by IP-Adapter.

Human evaluation. To assess the perceptual quality of EchoGen, we conduct a human evaluation study against strong baselines that span multiple categories of subject-driven methods. We focus on three criteria: text alignment, subject fidelity, and photorealism. The images are generated conditioned on the reference images and prompts sampled from DreamBench (Ruiz et al., 2023) and DreamBench++ (Peng et al., 2024) benchmarks without any cherry-picking, and for each criterion, participants select their preferred generated image among the outputs from five methods. We collect 450 responses from 25 participants, all with expertise in generative models, and report preference ratios in Table 2. The results show that EchoGen is preferred for subject fidelity and photorealism, surpassing all the diffusion-based contemporary baselines on these criteria. For text alignment, EchoGen performs on par with EasyControl (Zhang et al., 2025) and exhibits a clear advantage over the other compared methods.

Sampling Latency Analysis. We conduct a thorough analysis to evaluate the performance-latency trade-offs across all methods. Specifically, we evaluate diffusion-based methods with varying numbers of denoising steps and report their performance versus sampling latency in Figure 5. For the diffusion baselines, increasing the number of denoising steps improves subject fidelity (as measured by DINO, CLIP-I scores) up to a saturation point. In contrast, the text alignment (CLIP-T) score converges much earlier. Our model consistently offers a better trade-off, achieving superior performance with significantly lower sampling latency than the diffusion-based baselines. This confirms the inherent efficiency and effectiveness of our approach.

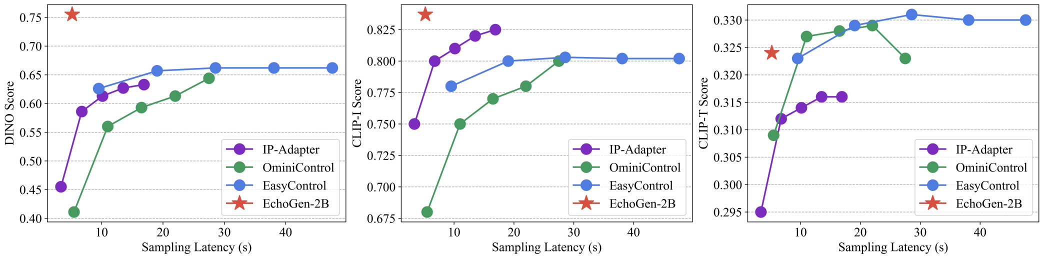


Figure 5: **Performance *v.s.* sampling latency comparison among our EchoGen and baselines.**

Enc.	DINO↑	CLIP-I↑	CLIP-T↑
SigLIP-2	0.438	0.720	0.320
FLUX.1-dev	0.433	0.706	0.320
DINOv2	**0.670**	**0.798**	**0.322**

Table 4: **Significance of fine-grained semantic injection**. “Enc.” denotes the encoder type.

Exp.	DINO↑	CLIP-I↑	CLIP-T↑
w/o prefix	0.632	0.788	**0.328**
w prefix	**0.670**	**0.798**	0.322

Table 5: **Ablation study on incorporating the global semantic features of reference images**.

The detailed component-wise sampling latency of our EchoGen framework is provided in Table 3. The results confirm that the framework's overall efficiency is not limited by auxiliary components such as Grounding-DINO. EchoGen maintains a significant speed advantage over diffusion-based methods, even with the inclusion of the optional Qwen2.5-VL model. Although this model is employed during training to automate subject identification for the GroundingDINO segmentation model, it is not required during inference. Instead, users can provide a descriptive text prompt (akin to the DreamBench format) for specifying the subject.

Model Component	Sampling Latency (s)
Grounding-DINO	0.24
Semantic encoder	0.008
Content encoder	0.025
Infinity generator	4.95
Qwen2.5-VL(Optional)	1.13
EchoGen (w/o Qwen2.5-VL)	5.22
EchoGen (w/ Qwen2.5-VL)	6.35

Table 3: **Per-component sampling latency measured on a single H20 GPU.**

5.3 Ablation Studies

We conduct a series of ablation studies to verify the effect of each component in EchoGen. Owing to computational constraints, all experiments are performed on EchoGen-0.1B with the same training settings to ensure fair ablation studies.

Significance of fine-grained semantic information injection. Fine-grained semantic conditional information is critical as it provides guidance for establishing the structure, enabling the model to synthesize stylistically and structurally coherent features consistent with the subject. Conversely, we argue that overly coarse-grained semantic features may fail to provide sufficient guidance for generating visually consistent echoes. To validate the importance of incorporating fine-grained semantic information, we conducted an ablation study with three distinct feature types independently injected via cross-attention:(1) coarse-grained semantic identity from SigLIP-2 (Tschannen et al., 2025), (2) fine-grained semantic features from DINOv2 and (3) FLUX.1-dev VAE features, which lack enough semantic information. Table 4 demonstrates that the fine-grained semantic DINOv2 features are the most suitable to represent the echo information in this task, as evidenced by all criteria. The failure of the SigLIP-2 and FLUX.1-dev VAE features can be attributed to their respective limitations: the former relies on features that are too coarse to guide subject generation, while the latter lacks semantic information.

Incorporating the global semantic information. To provide stronger guidance when constructing the global structure in the generation process, we prefix the global semantic C token extracted by the DINOv2 semantic encoder and incorporate this token into the model via Adaptive LayerNorm. As demonstrated in Table 5, compared with solely relying on injecting fine-grained semantic features via cross-attention, the introduction of the global semantic token yields a substantial performance gain in consistency of subject characteristics, validating the effectiveness of incorporating global semantic information.

Distinct semantic feature injection strategies. We explore the most effective method to guide the synthesis process conditioned on the semantic features of reference images. Table 6 presents an analysis comparing two distinct feature injection modules: multi-modal attention and cross-attention. Our results indicate that while the multi-modal module achieves slightly better alignment with text prompts, the cross-attention mechanism yields significantly superior subject fidelity, as evidenced by a notably higher DINO score. Based on this, we opted to utilize cross-attention for injecting the semantic features in all subsequent experiments, rather than the multi-modal attention.

Module	DINO↑	CLIP-I↑	CLIP-T↑
MM-Attn	0.646	0.792	**0.325**
Cross-Attn	**0.670**	**0.798**	0.322

Table 6: **Different methods for incorporating semantic features**.

Enhancing subject fidelity with detailed content features. Considering the absence of local details in the semantic features of the reference images, we incorporate a secondary pathway that injects localized content features of the subject. These features, extracted by the FLUX.1-dev VAE, are used to guide the synthesis of the fine-grained local details of the subject. The ablation study detailed in Table 7 demonstrates that employing a multi-modal attention mechanism to infuse these content features substantially enhances the subject-fidelity of the generated samples, yielding a significant increase in the CLIP-I metric.

Exp.	DINO↑	CLIP-I↑	CLIP-T↑
baseline	0.670	0.798	**0.322**
+Cross-Attn	0.667	0.803	0.318
+MM-Attn	**0.672**	**0.806**	0.321

Table 7: **Impact of injecting subject details.**

Qualitative analysis of the effect of semantic and content feature injection. We further qualitatively dissect the effect of each feature component to validate our design. As shown in Figure 6, starting from the base Infinity backbone, introducing semantic features extracted by DINOv2 enables the generator to synthesize subjects that faithfully preserve the reference subject's structure and style. Moreover, further incorporating content features from the FLUX.1-dev VAE significantly enhances EchoGen's capability to render fine-grained, coherent details (*e.g.*, the facial features of the robot toy and the fluffy dog, as well as the material and color of the shoe uppers). These qualitative results confirm the effectiveness of our dual-path injection design, where semantic and content features play distinct yet complementary roles.

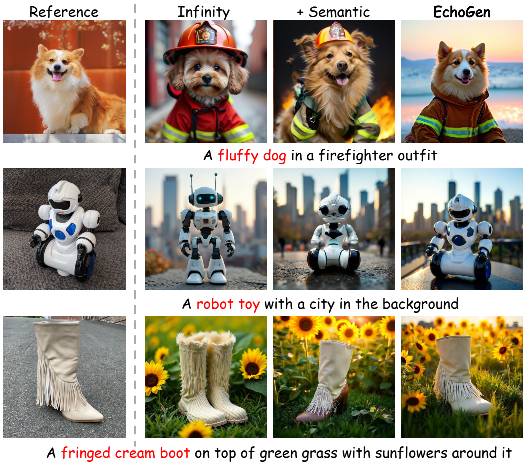

Reference
Infinity
+ Semantic
EchoGen
A fluffy dog in a firefighter outfit
A robot toy with a city in the background
A fringed cream boot on top of green grass with sunflowers around it

Figure 6: **Qualitative analysis of the effect of semantic and content feature injection.**

6 Conclusion

This paper presents EchoGen, a novel framework for efficient, feed-forward subject-driven image synthesis based on a visual autoregressive paradigm, aiming to inherit the properties of high-quality generation and fast inference speed. Central to our methodology is a dual-path injection mechanism, meticulously designed to integrate both the semantic attributes and the precise textural details of reference images. Comprehensive evaluations corroborate the superiority of our design, revealing that EchoGen achieves generative performance on par with leading diffusion models while exhibiting substantially lower sampling latency. By pioneering a feed-forward, autoregressive solution for subject-driven synthesis, this research charts a new trajectory for the future development and application of visual autoregressive generative models.

References

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, pp. 4055–4075. PMLR, 2023.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=XSEBx0iSjFQ`.

Jiwoo Chung, Sangeek Hyun, Hyunjun Kim, Eunseo Koh, MinKyu Lee, and Jae-Pil Heo. Fine-tuning visual autoregressive models for subject-driven generation. *arXiv preprint arXiv:2504.02612*, 2025.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

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

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=jQP5o1VAVc`.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL `https://openreview.net/forum?id=qw8AKxfYbI`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

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

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024b.

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. In *International Conference on Learning Representations*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=8ishA3LxN8`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=he6mX9LTyE`.

Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. λ-ECLIPSE: Multi-concept personalized text-to-image diffusion models by leveraging CLIP latent space. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=7q5UewlAdM`.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=di52zR8xgf`.

Senthil Purushwalkam, Akash Gokul, Shafiq Joty, and Nikhil Naik. Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. In *European Conference on Computer Vision*, pp. 252–269. Springer, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the pixelCNN with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=BJrFC6ceg`.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7986–7996, 2025.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2333–2343, 2025.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

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

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PJqP0wyQek.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

Ziyu Yao, Jialin Li, Yifeng Zhou, Yong Liu, Xi Jiang, Chengjie Wang, Feng Zheng, Yuexian Zou, and Lei Li. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=pfNyExj7z2.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=AFDcYJKhND. Featured Certification.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6786–6795, 2024.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025.

7 Appendix

7.1 Dataset

Following common practices like OminiControl (Tan et al., 2025) and EasyControl (Zhang et al., 2025), we train EchoGen on the synthetic Subjects200K dataset. The Subjects200K dataset[1], introduced by (Tan et al., 2025), comprises approximately 256,000 data triplets, each consisting of a reference image, a textual prompt, and a corresponding generated sample. This dataset was specifically established for subject-driven generative tasks. The curation process began by leveraging ChatGPT-4o to generate over 30,000 diverse subject descriptions, each description portraying the same subject across multiple scenes. Subsequently, these descriptions were reformulated into structured prompts, each detailing a single subject in two distinct scenes. These prompts were then used as input for the FLUX.1-dev model to synthesize image pairs. In the final stage, GPT-4o was employed to filter the resulting pairs, validating both subject consistency and overall image quality. All images within the dataset feature a resolution exceeding 500 pixels, ensuring sufficient detail for model training.

7.2 Pseudo-Code of the EchoGen Block

We provide a PyTorch-style pseudocode for our EchoGen Block in Algorithm 1 to facilitate reproducibility and clarity.

7.3 Implementation Details

7.3.1 Training Details

Data Preprocssing. All training images undergo a standardized pre-processing pipeline. Initially, images are resized so that their shorter edge matches the model's target resolution, which is 256 pixels for EchoGen-0.1B and 1024 pixels for EchoGen-2B, followed by a central crop to achieve a square aspect ratio.

To maintain data quality for high-resolution image generation within the EchoGen-2B model, we circumvent the quality degradation induced by naive bilinear upsampling. Instead of using simple bilinear scaling on images smaller than 1024 pixels, we integrate a super-resolution step. Specifically, we leverage the PiSA-SR model (Sun et al., 2025) to upscale these images, a method chosen to preserve fine-grained textures and prevent the introduction of common interpolation artifacts. This ensures that the model is trained exclusively on high-quality and high-resolution exemplars.

Training Hyper-parameters. We follow the Infinity-2B standard training recipe (Han et al., 2025), and the detailed hyperparameter configurations used to train our EchoGen are provided in Table 8. To mitigate error accumulation, as mentioned in Infinity, we employ bitwise self-correction by randomly flipping bits in the input sequence with a probability of 0.3. To improve robustness against variations in instruction length, prompts are randomly truncated to a single sentence with a probability of 0.5 during training. The EchoGen models are trained using 32 H20 GPUs, requiring 2 weeks for the longest schedule (training our EchoGen-2B model for 20 epochs).

7.3.2 Evaluation Details

For our quantitative evaluation, we utilize the DreamBench dataset (Ruiz et al., 2023). The dataset comprises 30 distinct subjects, categorized into 9 animate pets (cats and dogs) and 21 diverse inanimate objects (*e.g.*, toys, sunglasses, backpacks). Each subject is associated with 25 textual prompts specifically designed to test the model's abilities in recontextualization, property modification, and accessorization. Our data preparation protocol is adapted from (Pan et al., 2024), which involves selecting a single reference image per subject and augmenting its subject identity phrase with descriptive keywords. The correspondence between the DreamBench dataset directory name and the augmented subject description is summarized as follows:

- backpack, backpack

[1] `https://huggingface.co/datasets/Yuanshi/Subjects200K`

```
class EchoGenBlock(nn.Module)
   def __init__(dim, mask):
      # Multi-modal attention QKV projectors for the image token sequence
      self.qkv_mm = nn.Linear(dim, 3*dim)
      # Multi-modal attention QKV projectors for the detailed content feature
      self.qkv_mm_c = nn.Linear(dim, 3*dim)
      self.mask = mask

      # Cross attention query projectors for the image token sequence
      self.q_ca = nn.Linear(dim, dim)
      # Cross attention KV projectors for the semantic feature
      self.kv_ca_s = nn.Linear(dim, 2*dim)
      # Cross attention KV projectors for the text embedding
      self.kv_ca_t = nn.Linear(dim, 2*dim)

      # FFN for the image token sequence
      self.ffn = MLP(dim)
      # FFN for the detailed content feature
      self.ffn_c = MLP(dim)

   def forward(self, x, cc, cs, ct):
      # Multi-modal attention
      q, k, v = self.qkv_mm(x)
      qc, kc, vc = self.qkv_mm(cc)

      q = torch.concat((q, qc))
      k = torch.concat((k, kc))
      v = torch.concat((v, vc))
      x, cc = attention(q, k, v, self.mask)

      # Cross attention
      q = self.q_ca(x)
      ks, vs = self.kv_ca_s(cs)
      kt, vt = self.kv_ca_t(ct)

      k = torch.concat((ks, kt))
      v = torch.concat((vs, vt))
      x = attention(q, k, v, mask=None)

      # Feed-forward network
      x = self.ffn(x)
      cc = self.ffn_c(cc)

      return x, cc, cs, ct
```

Pseudo-code illustrating the EchoGen Block. Here, x denotes the image token sequence, and the generation process is conditioned on the semantic feature c_s and detailed content feature c_c extracted from the reference image, along with the text embedding c_t.

- backpack_dog, dog shaped backpack
- bear_plushie, bear plushie
- can, 'Transatlantic IPA' can
- candle, jar candle
- cat, tabby cat
- cat2, grey cat
- clock, number '3' clock
- colorful_sneaker, colorful sneaker
- dog1, fluffy dog
- dog2, fluffy dog
- dog3, curly-haired dog
- dog5, long-haired dog
- dog6, puppy
- dog7, dog
- dog8, dog
- duck_toy, duck toy

Config	value
Bitwise Self-correction Flip Ratio	0.3
Bitwise Self-correction Apply Layers	13
Dynamic Truncate Prompt Ratio	0.5
Infinity Image Encoder Channel	16(0.1B) / 32(2B)
Text Encoder	Flan-t5-xl
Text Embedding Channels	2048
Maximum Text Tokens Length	512
Semantic Image Encoder	DINO-v2-Base
Semantic Feature Channels	768
Semantic Downsample ratio	14
Content Image Encoder	FLUX.1-dev VAE
Content Feature Channels	16
Content Downsample ratio	8
Reweight Loss by Scale	True
Gradient clipping by norm	5.0
Optimizer	Adamw
Beta1	0.9
Beta2	0.97
Decay	0
Base Learning rate	3e-5
Multi-Modal Modules Learning rate	3e-6
Learning rate warmup iterations	0
Training epochs	20
Total Batchsize	128
GPU	H20

Table 8: **Detailed hyper-parameters for training our EchoGen**.

- fancy_boot, fringed cream boot
- grey_sloth_plushie, grey sloth plushie
- monster_toy, monster toy
- pink_sunglasses, sunglasses
- poop_emoji, poop-emoji shaped toy
- rc_car, car toy
- red_cartoon, cartoon character
- robot_toy, robot toy
- shiny_sneaker, sneaker
- teapot, clay teapot
- vase, tall vase
- wolf_plushie, wolf plushie

To augment the diversity and rigor of our human evaluation, we incorporate a curated set of instances from the DreamBench++ benchmark. DreamBench++ includes 150 subjects, each paired with nine prompts.

7.4 More ablation studies

In this section, we present additional ablation studies to analyze the individual components of EchoGen. These ablation studies are also conducted based on EchoGen-0.1B model with fair training settings.

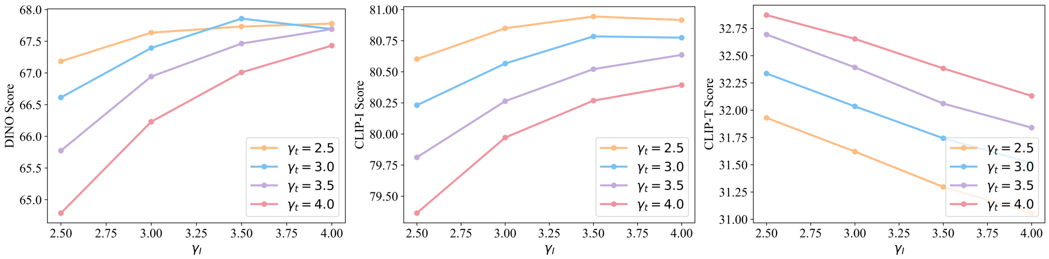


Figure 7: **Visualization of the effect of classifier-free guidance scale coefficient**.

Importance of injecting global semantic information. Injecting global semantic information serves as a prepended condition, to ensure global structural coherence during generation. Our ablation study in Table 5 confirms the significant benefits of incorporating global semantic features. Moreover, we conduct a targeted experiment comparing the injection of global semantic versus content features to the Image Token. As shown in Table 9, the results clearly indicate that prepending semantic features rather than content information into the image token significantly enhances the subject fidelity. This confirms that our choice to inject global semantic guidance into image tokens is both effective and well-justified.

Exp.	DINO↑	CLIP-I↑	CLIP-T↑
Content	0.663	0.795	**0.322**
Semantic	**0.672**	**0.806**	0.321

Table 9: **Analysis of different information types prepended to the Image Token.** Content represents prepending the global content feature, while Semantic denotes prepending the global semantic feature.

Subject Segmentation. To mitigate the influence of irrelevant background noise and focus on the primary subject, we leverage the Qwen2.5-VL vision language model (Bai et al., 2025) and the GroundingDINO segmentation model (Liu et al., 2024) to segment the subject from the reference image. We conducted an ablation study, detailed in Table 10, to validate the efficacy of the echo segmentation protocol. The results confirm that the introduction of subject segmentation significantly enhances the generation performance, which is observed in the preservation of subject features, and demonstrate that isolating the main subject is critical to producing more accurate and faithful outputs.

Exp.	DINO↑	CLIP-I↑	CLIP-T↑
w/o SS	0.663	0.796	0.321
w/ SS	**0.672**	**0.806**	0.321

Table 10: **Enhancement by subject segmentation (denoted by SS) to mitigate background noise**.

To further analyze the sensitivity of our method to the quality of segmentation, we conduct an ablation study on about EchoGen's robustness to segmentation quality during inference. Specifically, to simulate segmentation imperfections, we design three variants to simulate disturbances and compare with employing subject segmentation without imperfection during inference: 1. Enlarging Bounding Box: enlarging the subject's bounding box by 10%; 2. Shifting Bounding Box: shifting the bounding box by 10%; 3. No Segmentation: completely removing the subject segmentation step. As shown in Table 11, our model exhibits remarkable robustness as its performance degrades only slightly under these disturbances. Moreover, our model still produces strong results even without any segmentation, demonstrating its powerful generalization capability. In summary, EchoGen is highly robust to imperfect segmentation.

Exp.	DINO↑	CLIP-I↑	CLIP-T↑
w/o SS	0.737	0.829	**0.324**
Shift	0.739	0.833	0.321
Enlarge	0.735	0.831	0.321
w/ SS	**0.755**	**0.837**	**0.324**

Table 11: **Analysis of the sensitivity of our method to the quality of the segmentation.** SS denotes the subject segmentation; Enlarge and Shift denote enlarging and shifting bounding box, respectively.

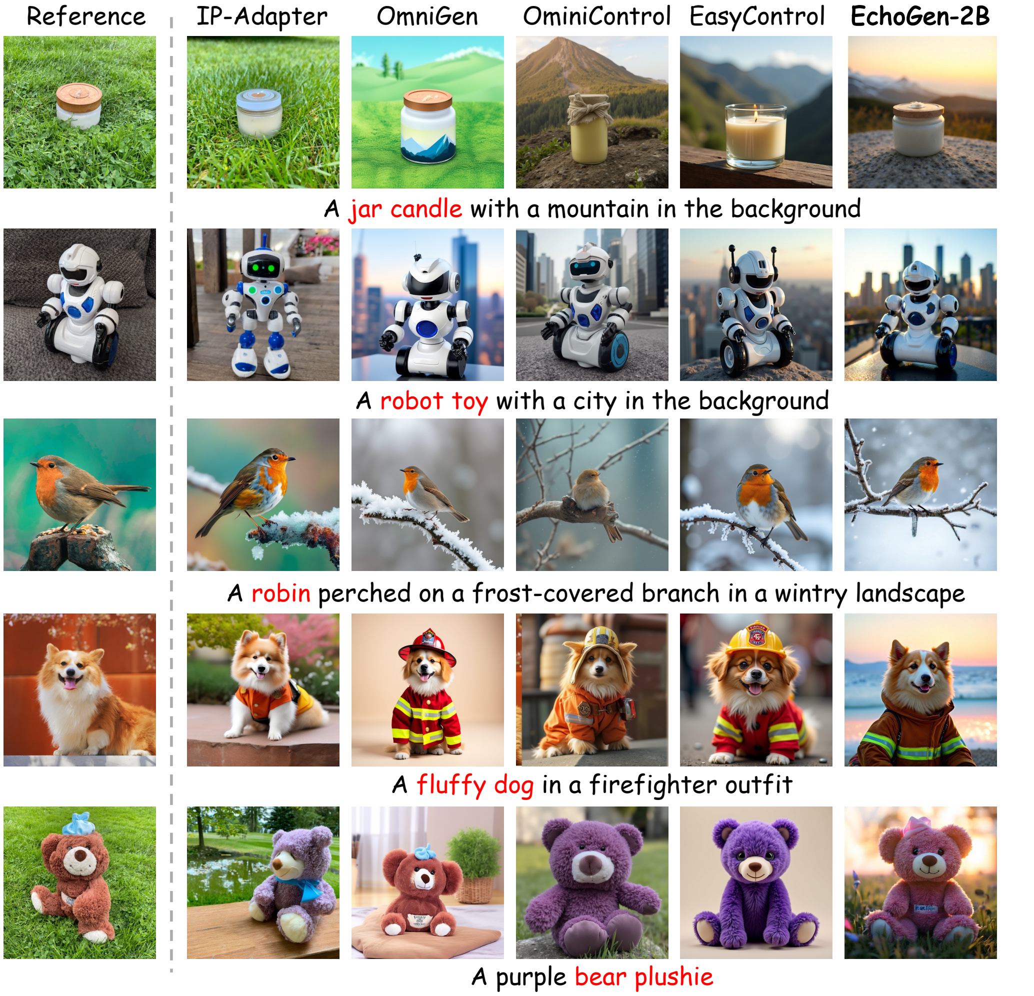

Reference
IP-Adapter
OmniGen
OminiControl
EasyControl
EchoGen-2B
A jar candle with a mountain in the background
A robot toy with a city in the background
A robin perched on a frost-covered branch in a wintry landscape
A fluffy dog in a firefighter outfit
A purple bear plushie

Figure 8: **Additional qualitative comparisons between our EchoGen model and competing methods.**

Subject-text classifier-free guidance. As detailed in Figure 7, our experiments reveal a clear trade-off governed by the CFG hyperparameters within a proper scope. As the subject guidance weight γ_I increases, subject fidelity improves, as indicated by higher CLIP-I and DINO scores. Conversely, this gain is accompanied by reduced text alignment, reflected in lower CLIP-T. The inverse relationship is observed when increasing the text condition scaling coefficient γ_t. This empirical result demonstrates the efficacy and flexibility of our CFG design, enabling users to dynamically adjust the balance between preserving reference image features and adhering to the text prompt.

7.5 More visualization results

We further showcase additional qualitative results on DreamBench in Figure 8. Moreover, we provide additional visual results from the EchoGen-2B model on real-world subject personalization in Figure 9. These results demonstrate that, with training exclusively on the large-scale high-quality synthetic dataset Subjects200K, our model exhibits strong generalization to real-world scenarios, including the generation of live animals and diverse objects under complex conditions. EchoGen-2B consistently maintains high subject fidelity and strong text alignment during these real-world

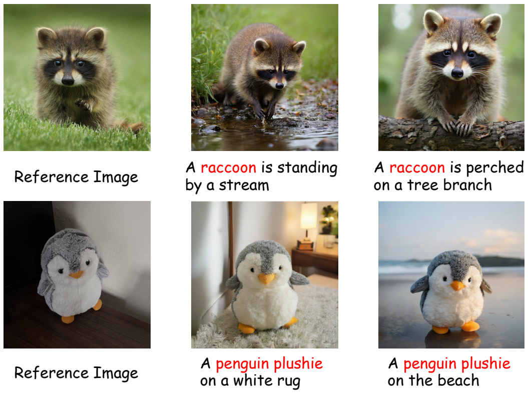
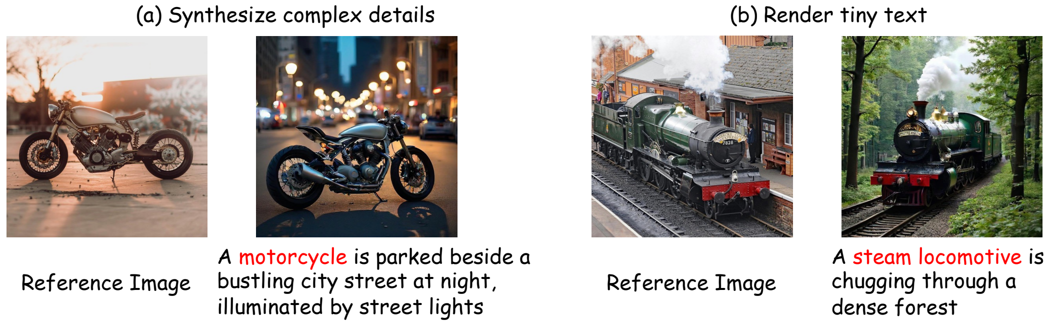

Reference Image
A raccoon is standing by a stream
A raccoon is perched on a tree branch
Reference Image
A penguin plushie on a white rug
A penguin plushie on the beach

Figure 9: **More visualization of EchoGen-2B on real-world subject personalization.**



(a) Synthesize complex details

Reference Image

A motorcycle is parked beside a bustling city street at night, illuminated by street lights

(b) Render tiny text

Reference Image

A steam locomotive is chugging through a dense forest

Figure 10: **Failure cases generated by EchoGen.**

personalization tasks, demonstrating the effectiveness of our training strategy and the proposed dual-path semantic-content injection design.

7.6 Limitation & Failure case analysis

Our EchoGen takes a new step toward VAR-based feed-forward subject-driven generation to inherit the strong capability of next-scale prediction and bidirectional modeling within scales. However, we know that the feed-forward subject-driven image generation is highly dependent on the capability of base models. The performance of our EchoGen models is fundamentally dependent upon the capability of the base models Infinity-0.1B and Infinity-2B. The Infinity-2B architecture still exhibits a performance gap compared to state-of-the-art generation models such as Stable-Diffusion 3 and FLUX, particularly in generating high-fidelity details. This inherited constraint limits EchoGen's efficacy in resolving fine-grained features, such as the faithful rendering of facial characteristics, the synthesis of coherent text, and the reproduction of intricate material textures. Due to significant GPU computational and temporal constraints, our experiments are confined to these specific backbones, precluding an empirical investigation of larger models such as Infinity-8B. We hypothesize that migrating the EchoGen architecture to a more potent VAR foundation model would unlock substantial performance gains.

Additionally, the DINOv2 vision encoder operates on relatively low-resolution inputs (*e.g.*, 224×224), which limits its ability to capture fine-grained appearance cues and tiny textual elements. We believe seeking an effective high-resolution semantic encoder presents a promising avenue for further improvement in complex applications.

Due to the aforementioned limitations, as illustrated in Figure 10, the model exhibits reduced reliability on subjects that have highly intricate structures or in scenarios requiring precise text rendering. We will continue to investigate and address these challenges in future work.

7.7 The usage of Large Language Model

We utilized the large language models Qwen-3 and GPT-5 to improve the clarity, grammar, and formal tone of the writing in the method and experiment sections. Nevertheless, all technical content, such as conceptual formulas and experiments remain our own; the large language models were used solely as tools for linguistic enhancement and stylistic polishing.