UniGen-AR: Unifying Visual Generation with Auto-Regressive Modeling

Zhipeng Bao[1] Zhen Zhu[2] Nupur Kumari[1]
Anurag Bagchi[1] Yu-Xiong Wang[1] Pavel Tokmakov[3†] Martial Hebert[1†]
[1]Carnegie Mellon University [2]University of Illinois Urbana-Champaign [3]Toyota Research Institute

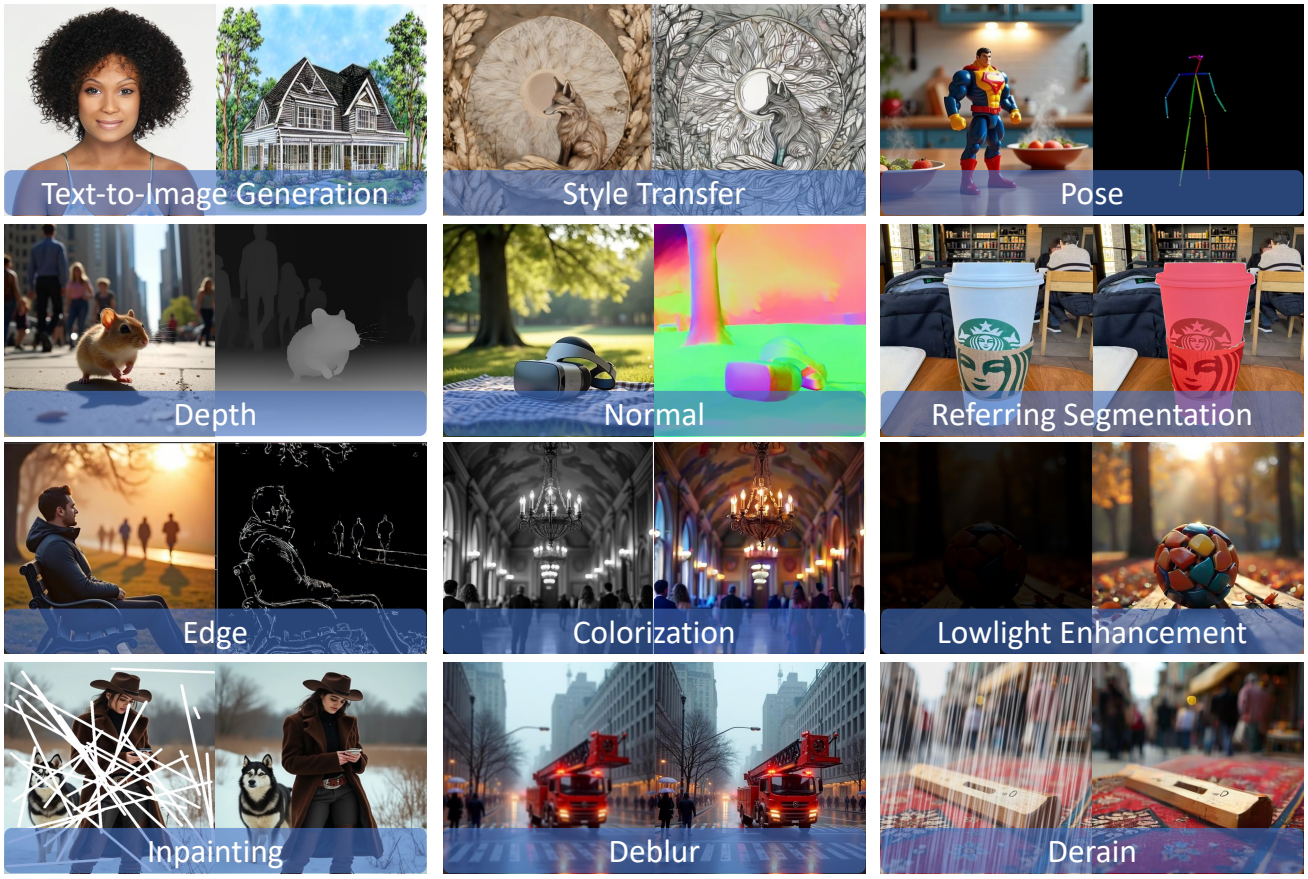

Figure 1. **UniGen-AR: A Single Model for Unified Visual Generation.** Our framework jointly handles 12 diverse tasks, spanning text-to-image synthesis, restoration, and classical perception. All outputs are generated by a single, MLLM-conditioned auto-regressive backbone, using a unified prompting interface and a single set of model weights without any task-specific heads.

Abstract

Modern computer vision pipelines remain fragmented, with tasks like text-to-image generation, editing, restoration, and classical perception handled by separate models. We study Unified Visual Generation (UVG), a setting where a single model jointly addresses all these tasks. While diffusion-based frameworks currently dominate UVG due to their high quality and controllability, their iterative sampling process incurs significant inference latency and can lead to unintended global edits. To address these limitations, we propose UniGen-AR, a framework that pairs a general-purpose multi-modal language model (MLLM) with an efficient visual auto-regressive (VAR) decoder. This design retains the flexibility of MLLM-based conditioning while leveraging the sampling efficiency and latent unification properties of VAR models. In our framework, the MLLM encodes free-form instructions and control signals into a unified sequence, which guides the VAR decoder to generate image-valued outputs for 12 tasks spanning three families. Empirically, UniGen-AR achieves up to $5\times$ lower inference latency than diffusion-based baselines while maintaining or improving output quality. Our ablations further reveal that VQ-VAE tokenizer design, par-

ticularly codebook size and hierarchy, is a critical factor for VAR scalability. These results establish visual auto-regressive modeling as a compelling and efficient backbone for unified visual generation.

1. Introduction

Modern visual generation pipelines remain fragmented: text-to-image synthesis, local and global editing, restoration, and classical vision tasks – such as depth and surface normal estimation and semantic segmentation – are typically handled by separate models or specialized heads [2, 41, 53, 83, 84]. In this work, we study *Unified Visual Generation* (UVG) [27, 37, 74], a setting in which a single model produces diverse *image-valued outputs* under a shared interface. UVG promises shared representations, consistent controllability, and operational simplicity. However, scaling a single model to support many output types and control modalities is non-trivial: as the catalog of tasks and outputs grows, the model must learn to handle heterogeneous visual domains and control signals while maintaining coherent behavior across all their combinations. This imposes a combinatorial burden and severe capacity and optimization challenges, making scalability the central obstacle for effective UVG.

To tackle this scaling challenge, current approaches to UVG are dominated by diffusion-based architectures [25] paired with powerful multimodal encoders [3]. This bundled design directly addresses the combinatorial challenge: the MLLM acts as a universal interface to interpret heterogeneous control signals (*e.g.*, text, reference images), while the diffusion backbone learns a shared generative representation for the diverse image-valued outputs. These models, *e.g.*, Batifol et al. [5], Comanici et al. [10], Wu et al. [74], set a high standard for output quality and task coverage. However, they incur significant computational costs: iterative denoising, bundled with MLLM inference, yields significant inference latency, and the global nature of the denoising process often leads to unintended spurious edits [46]. These limitations motivate the search for alternative architectures that retain rich conditioning while offering a better **latency–quality trade-off**.

To this end, we revisit *auto-regressive* (AR) modeling as an alternative backbone for UVG. While naïve AR over coarse image tokens often struggles with fine-grained detail, recent work on *visual auto-regressive* (VAR) modeling demonstrates that *next-scale prediction* over discrete visual tokens can achieve high-fidelity synthesis, stable likelihood-based training, and efficient sampling [64]. In the context of UVG, VAR presents two key advantages: (i) *latent unification* – similar to diffusion, VAR decoders can produce both natural images and structured predictions, like depth maps, within a shared token space; and (ii) *sampling efficiency* – AR models typically require significantly fewer steps than diffusion to achieve comparable perceptual quality [21, 64].

The VAR backbone provides sampling efficiency, but to achieve the rich, instruction-based controllability required for UVG [37, 74], it must be paired with a powerful multimodal front-end. We therefore propose **UniGen-AR**, a framework that couples a general-purpose multi-modal language model (MLLM) with a visual auto-regressive decoder. This architecture is designed to retain diffusion-style flexibility while benefiting from AR efficiency. In our framework, the MLLM encodes free-form instructions and diverse control signals (*e.g.*, text, reference images) into a unified conditioning sequence. The auto-regressive decoder then predicts discrete visual tokens conditioned on this sequence, which are decoded into the final image or dense map output via a VQ-VAE [20].

To evaluate our proposed framework, we instantiate UniGen-AR by re-purposing a powerful pre-trained *text-to-image* VAR model [21] for the full UVG setting. We train a single backbone jointly across *12 tasks* spanning three families: text-to-image generation, classic perception, and restoration. Typical visualizations are shown in Figure 1. Training is performed with a unified likelihood objective under the consistent MLLM-conditioned interface. Empirically, UniGen-AR achieves strong performance across all 12 tasks, outperforming prior AR-based systems and exhibiting especially strong results on restoration and classical perception. Compared to diffusion-based UVG models under matched conditioning, UniGen-AR presents a favorable latency–quality Pareto frontier, achieving up to $\sim 5\times$ lower inference latency while maintaining or improving output quality on representative benchmarks. Ablation studies further reveal the design of the VQ-VAE tokenizer, particularly codebook size and hierarchy, as a critical factor influencing performance at scale.

Our contributions are summarized as follows:

- We present, to our knowledge, the first framework that scales visual auto-regressive modeling to the full UVG setting, unifying open-ended synthesis, restoration, and visual perception within a single image-out backbone.
- We demonstrate a compelling latency–quality trade-off compared to diffusion-based systems, achieving consistent speedups while maintaining or improving performance.
- We identify and validate the importance of VQ-VAE tokenizer design, including codebook size and hierarchy, as a key driver of VAR scalability and effectiveness.
- We study the bidirectional connection between multimodal understanding and generation, showcasing how understanding-enhanced MLLM front-ends improve control and quality in image synthesis.

2. Related Work

Unified visual generation aims to support tasks such as text-to-image generation, editing, restoration, and perception within a single model. Early methods explored shared latent spaces using variational autoencoders (VAEs) and vision transformers [31, 43, 44]. More recent approaches have adopted diffusion-based models, pretrained on large-scale data, to address a wide range of generative tasks under a unified interface [7, 15, 22, 36, 38, 75, 77, 82]. As a follow-up, another line of work builds MLLM-mediated pipelines for controllable image generation [5, 37, 41, 49, 74]. For example, Qwen-Image [74], Step1X-Edit [41], and MetaQueries [49] pair powerful multimodal front-ends with diffusion decoders to support instruction-based rendering and precise editing. Flux-Kontext [5] and VisualCloze [37] focus specifically on in-context learning, enabling models to follow few-shot examples for visual tasks.

A second strand builds MLLM-mediated pipelines for controllable generation [5, 37, 41, 49, 74]. Among them, Qwen-Image [74], Step1X-Edit [41], and Metaquries [49] pair powerful multimodal front-ends with diffusion decoders for text rendering and precise editing. Flux-Kontext [5] and Visualcloze [37] specifically focus on in-context learning to enable models with the capability to learn from few-shot examples.

A smaller yet growing body of work investigates AR modeling for image generation [4, 18, 33, 58–60, 62]. These models typically adopt the standard "next-token prediction" formulation. In contrast, our work builds on the "next-scale prediction" paradigm introduced in visual autoregressive (VAR) models [64], which is better suited for UVG due to its coarse-to-fine decoding strategy.

Visual auto-regressive models factorize image generation into scale-wise predictions over discrete visual tokens, allowing coarse-to-fine decoding. The foundational work of Tian et al. [64] demonstrates favorable scaling laws and superior latency–quality trade-offs compared to diffusion models. Subsequent studies extend this formulation to conditional image generation in different domains beyond ImageNet [8, 35, 45, 51, 55, 68, 85].

Among them, two recent approaches adapt next-scale VAR to text-to-image generation [21, 66]. Both adopt cross-attention modules to inject text signals into the visual decoder, following a design similar to Stable Diffusion [53]. Switti [66] introduces a refined attention masking strategy that restricts each token to attend only to spatially local neighbors within the current scale, improving inference speed. Infinity [21] identifies large codebook sizes in VQ-VAE as key to achieving high-quality synthesis. A few recent works have begun extending VAR beyond text-to-image to support editing [46, 67, 70]. However, EditInfinity [67] and related methods [70] do not support direct instruction-guided UVG. Instead, they rely on indirect mechanisms such as modifying attention maps or text embeddings to steer edits. The concurrent VAREdit [46] adapts VAR models for editing, but focuses solely on this task and does not incorporate MLLMs or address broader UVG settings. In contrast, we present the first framework that combines next-scale VAR modeling with MLLM-based conditioning to support full-spectrum UVG tasks.

Unified models aim to handle both understanding (text-out) and generation (image-out) in a single architecture. One direction pursues tightly-coupled token-based models [28, 29, 60, 61, 63, 69, 76]. Chameleon [63] pioneered early-fusion, any-order modeling over text and image tokens in a single Transformer. Emu3 [69] extends this approach to support both understanding and generation over images and videos, using next-token prediction on discrete tokens. LaVIT [28] and its successors integrate LLMs with discrete visual tokenizers to perform both perception and synthesis tasks under a unified generative interface.

A second direction explores hybrid designs that decouple encoding and decoding while maintaining a single interface. These approaches pair MLLM front-ends with task-specific decoders [12, 26, 32, 48, 73, 81], often sharing a central autoregressive core to support both instruction following (text-out) and controllable generation or editing (image-out). This design balances task flexibility with operational simplicity. Our framework follows this hybrid philosophy. By coupling an MLLM front-end with a next-scale VAR decoder, we enable instruction-conditioned image generation across a wide task spectrum. Preliminary results also indicate that jointly fine-tuning the MLLM and visual decoder improves alignment between vision and language representations, suggesting a promising path toward fully unified multimodal models, which we leave for future exploration.

3. Method

This section presents the design of UniGen-AR (illustrated in Figure 2). We begin by reviewing VAR modeling as the core generative backbone. We then detail how we perform UVG based on existing T2I VAR models.

3.1. Preliminary: Visual Auto-Regressive Modeling

Vanilla VAR models. VAR [64] generates high-fidelity images by autoregressively predicting discrete visual tokens across multiple *spatial scales*. It operates in a latent space defined by a multi-scale vector-quantized tokenizer, typically implemented via a VQ-VAE [31, 65].

Given an image I, the encoder $\mathcal{E}$ produces K token maps:$R = \mathcal{E}(I) = (r_1, r_2, \ldots, r_K), \quad r_k \in [V]^{h_k \times w_k}$, where V is the codebook size, and spatial resolution increases with scale index k (*i.e.*, $h_1 w_1 \leq \cdots \leq h_K w_K$). The decoder $\mathcal{D}$ reconstructs the image via: $\hat{I} = \mathcal{D}(r_1, \ldots, r_K)$.

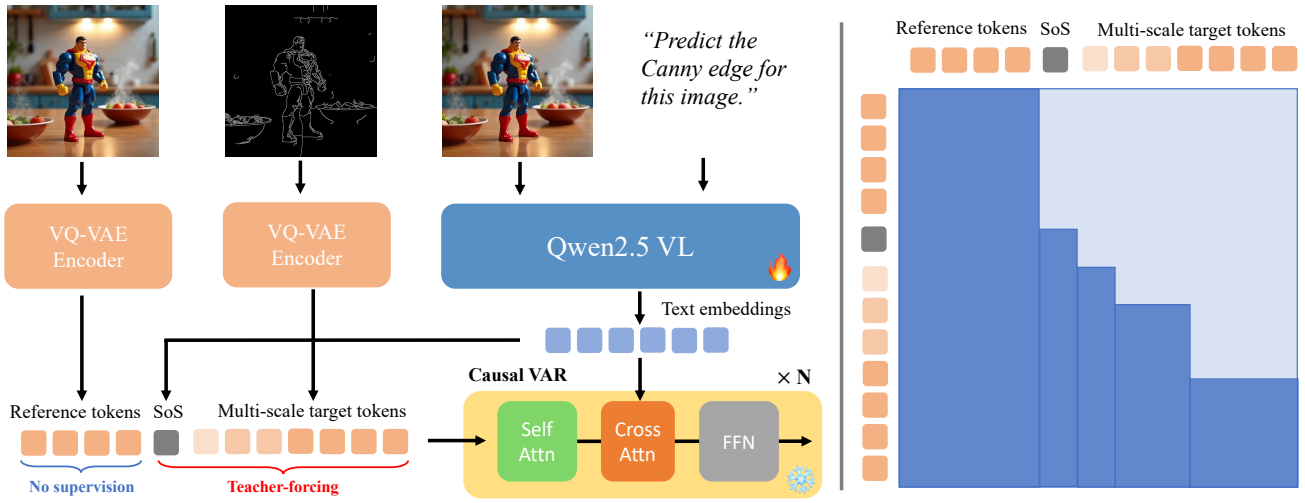

Figure 2. **Architecture of UniGen-AR.** Left: We extend an Infinity-style VAR backbone with a multimodal language model encoder to support Unified Visual Generation with reference images. The MLLM encodes the instruction and reference image, whose text embeddings seed a learnable [SoS] token and provide keys/values for cross-attention, while finest-scale reference tokens are prepended as a non-predictive context prefix; Right: A block-wise causal mask lets all target tokens attend to reference and text context while preserving the standard coarse-to-fine VAR schedule.

With the training multi-scale token sequence R, VAR defines an autoregressive factorization across scales:

$$p_\theta(R) = \prod_{k=1}^{K} p_\theta(r_k | r_{<k}), \tag{1}$$

where $r_{<k}$ denotes tokens from coarser scales. Each r_k is predicted in parallel, conditioned on $r_{<k}$ and scale-specific position embeddings. A block-wise causal mask ensures each token in r_k only attends to tokens in $r_{\leq k}$. Inference proceeds sequentially from $k = 1$ to K, with key-value caching for efficiency.

The model is trained with teacher forcing using the sum of cross-entropy losses over all token positions and scales:

$$\mathcal{L}_{\text{VAR}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \text{CE}\Big(p_\theta(\cdot \mid r_{<k}),\ r_k^{(i)}\Big), \tag{2}$$

where $n_k = h_k w_k$ is the number of tokens at scale k, and $r_k^{(i)}$ is the ground-truth token (from $\mathcal{E}$) at position i.

Text-conditioned next-scale prediction. To enable text-to-image generation, the VAR factorization in Equation (1) is extended to condition on a prompt t, encoded by a frozen language model $\psi(t)$:

$$p_\theta(R) = \prod_{k=1}^{K} p_\theta(r_k | {<k}, \psi(t)). \tag{3}$$

Following prior work [21, 66], the text embedding $\psi(t)$ is injected at each transformer layer via cross-attention. A learnable start-of-sequence ([SoS]) token, derived from a projection of $\psi(t)$, is prepended at the coarsest scale to bootstrap generation. Training remains unchanged, using block-wise causal masking and the loss in Equation (2).

Infinity [21], a state-of-the-art T2I VAR model serving as our backbone VAR model in our main experiments, replaces the standard categorical tokenizer with a *bitwise* tokenizer that encodes each visual token as a binary vector. Instead of predicting a single index, Infinity predicts the bits of this code, so that increasing the bit-width enlarges the effective vocabulary exponentially while only mildly growing the classifier head. This binary indexing enables extremely large visual vocabularies and thus higher-fidelity reconstruction and richer visual details. The bitwise design remains compatible with cross-entropy-style training, implemented as independent binary cross-entropy losses over bits. At inference time, Infinity conditions on the text prompt t and autoregressively generates the multi-scale residual token maps $(r_1, \ldots, r_K)$ with cached key–value attention states, following the VAR coarse-to-fine schedule.

3.2. From Text-to-Image to Unified Visual Generation

We use Infinity [21] as our T2I backbone and extend its architecture to UVG with reference images. To enable image-to-image transformations, the model must first understand the input images; we therefore introduce a multimodal encoder that processes reference images (and text) into a shared token space. Our overall design is illustrated in Figure 2.

Reference- and target-tokenization. Given a reference image I^{ref} and a target image I^{tar}, we first encode them

with the same multi-scale VQ-VAE encoder $\mathcal{E}$:

$$R^{\text{ref}} = \mathcal{E}(I^{\text{ref}}), \quad R^{\text{tar}} = \mathcal{E}(I^{\text{tar}}). \tag{4}$$

For the target image, we keep all scales $R^{\text{tar}} = (r_1^{\text{tar}}, \ldots, r_K^{\text{tar}})$ as in vanilla Infinity. For the reference image, we only retain the finest scale r_K^{ref}, and discard coarser scales:

$$R^{\text{ref}} = (r_K^{\text{ref}}), \quad r_K^{\text{ref}} \in [V]^{h_K \times w_K}. \tag{5}$$

The concurrent work, EditVAR [46], also demonstrates that finest-scale tokens are sufficient for UVG tasks. This design, prepending the finest reference tokens, leaves the original VAR scale schedule unchanged: the reference tokens are never traversed by the next-scale generation process and are excluded from the loss.

Multimodal encoder. We replace the Infinity's text encoder, T5 [9], with a multimodal language model ψ (Qwen2.5-VL [3]) to encode the input instruction t, together with the reference image: $Z_t = \phi(I^{\text{ref}}, t)$. Similar to prior efforts in the Diffusion regime [74], Z_t are the *pure text embeddings* from the last self-attention layer of the MLLM encoder to remove the redundancy. These text embeddings are used in two ways, following Infinity [21]: a learnable [SoS] token is obtained by projecting the text embedding and prepended at the coarsest scale to bootstrap generation, and the text embedding also serves as the key and value sequence for the cross-attention layers that modulate the visual tokens.

Unified token sequence and causal masking. We unify reference and target tokens into a single autoregressive sequence

$$\mathbf{s} = \left(r^{\text{ref}},\ [\text{SoS}],\ r_1^{\text{tar}}, \ldots, r_K^{\text{tar}}\right), \tag{6}$$

where the reference tokens are *prepended* before the [SoS] token. Conceptually, r^{ref} acts as a non-predictive context prefix, the [SoS] token marks the autoregressive start, and the multi-scale target tokens follow the standard VAR schedule.

We implement a block-wise causal mask $M \in \{0,1\}^{|\mathbf{s}|\times|\mathbf{s}|}$, visualized on the right side of Figure 2, that enforces:

- reference tokens are visible to all subsequent tokens ([SoS] and target tokens) but are never used as prediction targets;
- the [SoS] token and all target tokens respect the original next-scale causal ordering: tokens in scale k can attend to reference tokens, [SoS], and all tokens in $r_{<k}^{\text{tar}}$, but not to future scales.

Training and inference. Training follows teacher forcing as in Equation (2). At inference time, we perform iterative next-scale prediction as in standard VAR: we first sample r_1^{tar} conditioned on $\mathbf{r}^{\text{ref}}$ and t, then proceed to finer scales until r_K^{tar} is obtained. Due to the causal mask, tokens at each step can freely attend to the entire reference sequence and the text context while respecting the multi-scale ordering. When no reference image is provided, the same decoding procedure reduces to conventional T2I generation.

4. Experimental Evaluations

4.1. Experimental Setup

Training data. Following prior work [37, 72, 77], we train UniGen-AR using publicly available paired datasets. For text-to-image (T2I) generation, we use the LAION-COCO-Aesthetic subset [54, 77], containing approximately 4M images. For perception tasks, we adopt the Graph200K dataset from VisualCloze [37], which provides 200K images paired with annotations for depth estimation, surface normals, edge detection, and human pose estimation. For image restoration, we consider five tasks: deblurring, deraining, colorization, inpainting, and low-light enhancement. We follow the processing scheme of VisualCloze to generate the noisy version of these labels. For referring image segmentation, we use RefCOCO [80], which contributes roughly 320K (image, mask, text) triplets. For style transfer, we train on StyleBooth [23], containing around 11K images. In total, our training corpus comprises roughly 6M paired examples across 12 UVG tasks.

Implementation details. We initialize our system from the pretrained Infinity-2B model [21]. We replace its original T5 [9] text encoder with Qwen2.5-VL (3B) [3] for multimodal conditioning. Following [74], we extract only the textual embeddings from the MLLM and omit visual embeddings. Training proceeds in two stages: **Stage I (alignment).** We freeze the entire Qwen2.5-VL and Infinity backbones, except for the text-normalization layer, text-projection layer, and unconditional embeddings used for classifier-free guidance. This aligns the pretrained Infinity decoder with the new Qwen2.5-VL conditioning. In this stage, we only train the model with the T2I data. **Stage II (UVG training).** We jointly train the Infinity backbone components on the full mixture of 12 tasks. Each batch is either a pure T2I batch or a mixed batch drawn from all other tasks, with the T2I sampling probability set to 0.25. All outputs are generated at a fixed resolution of 512×512.

Stage I is trained for 2 epochs with an effective batch size of 256; Stage II is trained for 100K steps with an effective batch size of 128. We use AdamW [42] with learning rates of 1e-4 (Stage I) and 5e-6 (Stage II). Training requires approximately 3 days for Stage I and 7 days for Stage II on a single NVIDIA H100 node.

Evaluation benchmarks. We evaluate on three groups of tasks. **T2I generation:** GenEval benchmark [19], following Infinity [21]. **Perception tasks:** depth estimation and surface normals on NYUv2 [56]. **Restoration tasks:** low-light enhancement on LOL [71], deblurring on GoPro [47],

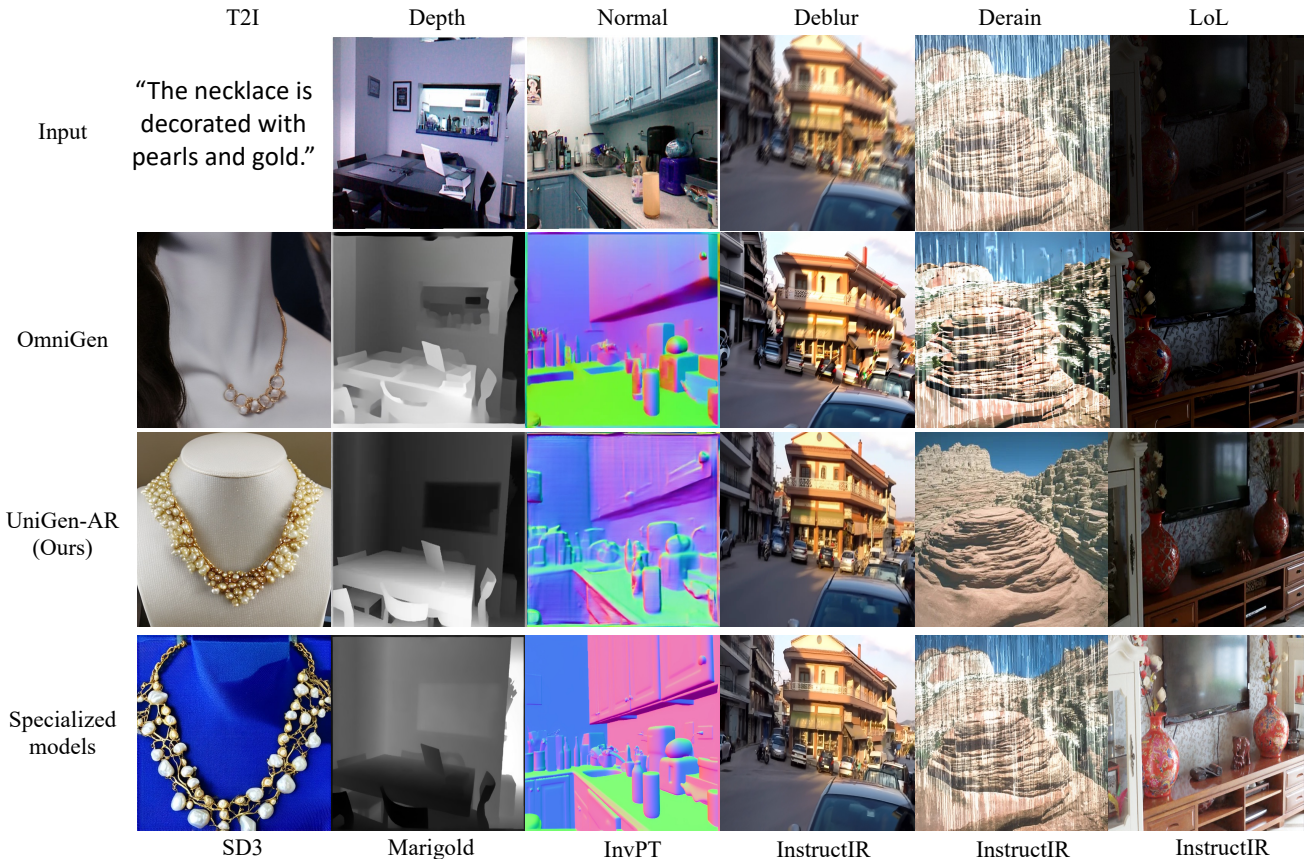

Figure 3. **Qualitative comparisons across UVG tasks.** UniGen-AR consistently produces higher-quality results than the UVG baseline OmniGen and achieves visually comparable or superior outputs to specialized models, notably outperforming InstructIR [11] on deraining.

Model	# Params	GenEval			
		Two Obj.	Position	Color Attr.	Overall
SDv1.5 [53]	0.9B	0.38	0.04	0.06	0.43
SDv2.1 [53]	0.9B	0.51	0.07	0.17	0.50
DALL-E 2 [52]	6.5B	0.66	0.10	0.19	0.52
DALL-E 3 [6]	-	-	-	-	0.67
SDXL [50]	2.6B	0.74	0.15	0.23	0.55
SD3 (d=21) [14]	2B	0.74	0.34	0.36	0.62
LlamaGen [57]	0.8B	0.34	0.07	0.04	0.32
Chameleon [63]	7B	-	-	-	0.39
Emu3 [69]	8.5B	0.81	**0.49**	0.45	0.66
Infinity [21]	2B	**0.85**	**0.49**	**0.57**	**0.73**
UniGen-AR (Ours)	2B	0.76	0.41	0.45	0.68

Table 1. **GenEval Text-to-Image Results.** Comparison with diffusion-based models (top) and autoregressive models (bottom). UniGen-AR achieves competitive performance while using only public data and supporting 12 unified visual generation tasks.

and deraining on Rain-13K [16].

Qualitative comparisons for all tasks appear in Figures 1 and 3, with additional examples provided in the supplementary. More details about our training data and implementation are also included in the supplementary.

4.2. Main results

We report results on T2I generation (Table 1), perception tasks, and image restoration (Table 2). Representative outputs are shown in Figure 3. For T2I, we compare against both diffusion-based and autoregressive models; for perception and restoration tasks, we compare with specialized task models and recent UVG systems.

Text-to-image generation. Table 1 shows that: (1) UniGen-AR achieves strong performance on GenEval, outperforming larger diffusion-based models such as DALL-E 2 [52], demonstrating that next-scale VAR remains competitive even in the unified setting. (2) Compared with the Infinity checkpoint, our model shows a mild drop in performance, likely due to Infinity's use of large-scale proprietary training data. (3) Despite using only public data, our model surpasses diffusion counterparts with similar model sizes, including SD3 (2B) [14]. This suggests that VAR-based backbones retain strong prior knowledge during finetuning and remain a compelling alternative to diffusion for controllable image generation.

Perception and image restoration tasks. Based on the results in Table 2, we offer the following observations: (1) UniGen-AR consistently outperforms the strongest AR-based UVG model X-Prompt [60] across all evaluated tasks, highlighting the advantage of coarse-to-fine refinement in next-scale prediction. (2) Compared with specialized task-

Model	NYUv2-Depth RMSE (↓)	NYUv2-Normal Mean Angle Err ↓	LOL-Lowlight PSNR (↑)	LOL-Lowlight SSIM (↑)	GoPro-Deblur PSNR (↑)	GoPro-Deblur SSIM (↑)	Rain100L-Derain PSNR (↑)	Rain100L-Derain SSIM (↑)
Depth Anything [78]	**0.206**	-	-	-	-	-	-	-
Marigold [30]	0.224	-	-	-	-	-	-	-
Bae *et al*. [1]	-	**14.90**	-	-	-	-	-	-
InvPT [79]	-	19.04	-	-	-	-	-	-
AirNet [34]	-	-	18.18	0.735	24.35	0.781	32.98	0.951
InstructIR [11]	-	-	**23.00**	**0.836**	**29.40**	**0.886**	**36.84**	**0.937**
InstructCV [17]	0.297	-	-	-	-	-	-	-
UnifiedIO [43]	0.387	-	-	-	-	-	-	-
OmniGen [77]	0.480	-	13.38	0.392	13.39	0.321	12.02	0.233
X-Prompt [60]	0.277	19.17	19.71	0.810	21.04	0.761	25.53	0.843
UniGen-AR (Ours)	**0.245**	**18.76**	**21.03**	**0.825**	**22.99**	**0.774**	**33.71**	**0.926**

Table 2. **Results on Perception and Restoration Tasks.** UniGen-AR significantly outperforms AR-based UVG prior work (X-Prompt) and demonstrates competitive performance on image restoration tasks.

Model	Two Obj.	Position	Color Attr.	Overall	inf. time (s/img)
SD3 (d=21)	0.74	0.34	0.36	0.62	3.18
Infinity	0.85	0.49	0.57	0.73	0.92
Qwen+SD3	0.68	0.36	0.29	0.52	5.23
Qwen+Infinity	0.75	0.45	0.43	0.64	1.05

Table 3. **Diffusion *vs*. VAR decoders.** We compare SD3 (diffusion) and Infinity (VAR) under identical finetuning settings. VAR provides better generation accuracy and is substantially faster at inference.

Model	LOL-Lowlight PSNR (↑)	GoPro-Deblur PSNR (↑)	Rain100L-Derain PSNR (↑)
Infinity + T5	20.10	22.17	29.32
Infinity + Qwen	**21.03**	**22.99**	**29.71**

Table 4. **Effect of multimodal encoder.** Replacing T5 with Qwen2.5-VL leads to significant improvements on all restoration tasks, indicating the benefit of multimodal grounding.

specific models, a performance gap remains—reflecting the inherent challenge of UVG, where a single model must master diverse, heterogeneous objectives. (3) Notably, our model showcases superior performance on the restoration tasks. In particular, for low-light enhancement and derain tasks, our model surpasses a dedicated restoration model (AirNet [34]), suggesting that the bitwise VQ-VAE used in Infinity provides a favorable structure for correcting token-level degradations.

Qualitative comparisons. Figure 3 illustrates that UniGen-AR produces higher-quality results than the UVG baseline OmniGen [77] across all evaluated tasks. Moreover, for most tasks, our outputs are visually comparable to those of specialized models. Notably, on the deraining task, UniGen-AR achieves cleaner rain removal than the state-of-the-art dedicated model InstructIR [11]. These qualitative results further highlight the strength of our MLLM–VAR architecture for UVG and suggest promising potential for real-world applications.

4.3. Ablation Study

Diffusion *vs*. VAR. We first compare the impact of the decoder architecture by replacing the Infinity VAR decoder with the SD3 diffusion decoder, while keeping all other training settings (*i.e*., data, steps, resolution) identical. For efficiency, all variants are finetuned for two epochs on the T2I subset only; therefore, absolute numbers differ from Table 1. Results are summarized in Table 3.

Both models exhibit performance drops relative to their original checkpoints, primarily due to the substantially smaller and purely public finetuning data. Nevertheless, under this controlled setting, the VAR-based Infinity decoder consistently outperforms the diffusion-based SD3 decoder across all GenEval categories. This highlights next-scale prediction as a robust alternative to diffusion when finetuned jointly with a multimodal encoder. Interestingly, integrating Qwen2.5-VL improves the spatial grounding capability of the SD3 variant – its Position score increases from 0.34 to 0.36 with poorer training data, showcasing the benefit of replacing a text-only encoder with an MLLM for UVG. Finally, the Infinity variants achieve approximately **5× faster** inference than their SD3 counterparts, underscoring the practical appeal of VAR for real-time or interactive generation workloads.

Choice of multimodal encoder. To evaluate the value of multimodal conditioning, we compare Qwen2.5-VL [3] with the original T5 [9] encoder used in Infinity. Both variants are trained on the same data and with the same schedule. Table 4 reports results on three image restoration tasks. The Qwen-based model achieves notably higher PSNR across all tasks. This suggests that sending both the reference image and the instruction prompt into an MLLM yields text embeddings that implicitly encode object- and region-level semantics, which are more informative than the purely linguistic embeddings produced by T5. As UVG tasks often require localized reasoning, this multimodal

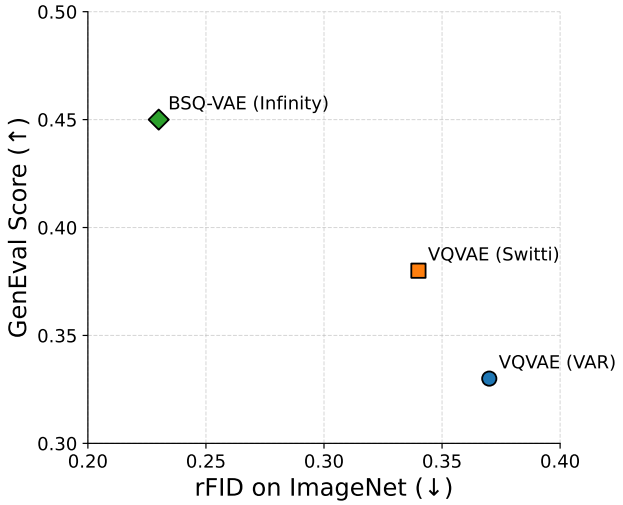

Figure 4. **Impact of visual tokenizers.** Better reconstruction fidelity leads to substantially improved generation quality, highlighting the tokenizer as a key design factor in VAR-based models.

grounding becomes particularly beneficial.

Impact of visual tokenizer. We further study the influence of the discrete visual tokenizer by training three variants using VQ-VAEs from VAR [64], Switti [66], and Infinity [21]. Different from the core experiment, here we finetune the checkpoint from the **vanilla VAR model** All models are trained in a single-stage T2I setting for two epochs. Figure 4 shows reconstruction FID (rFID) on ImageNet [13, 24] and GenEval [19] performance. We observe a strong inverse correlation between rFID and generation quality: tokenizers with lower reconstruction error yield higher GenEval scores. This underscores the visual tokenizer as a central bottleneck in VAR-style architectures. Improving token expressiveness and reconstruction fidelity remains a promising direction for advancing VAR generation.

4.4. Multimodel Understanding and Unified Visual Generation

Thus far, we have focused on how an MLLM can provide stronger conditioning signals for the VAR backbone. In this experiment, we take a slightly different perspective: treating UniGen-AR as a visual-output branch of an MLLM, we investigate whether improving the MLLM's understanding ability can, in turn, enhance its visual generation.

To control for data and isolate the effect, we exclusively reuse the same T2I subset employed for training the generator. Each T2I sample is repurposed into a VQA-style instance by converting the caption into an answer and assigning a fixed question template: *"Generate a caption for this image."* To increase linguistic diversity, we follow and pre-sample 50 paraphrased variants of this question via Qwen2.5-VL itself for training. During joint training, we finetune the last 10 layers of Qwen2.5-VL together with the Infinity decoder. The multimodal-understanding loss updates only the Qwen layers, whereas the UVG loss updates both Qwen and Infinity in a coupled manner. All variants are trained for two epochs on T2I data for fair comparison.

Setting	Two Obj.	Position	Color Attr.	Overall
Ours ((UVG only))	0.75	0.45	0.43	0.64
Ours (Joint MMU+UVG)	**0.82**	**0.47**	**0.46**	**0.69**

Table 5. **Multimodal understanding improves unified visual generation.** Jointly finetuning Qwen2.5-VL for multimodal understanding leads to consistent performance gains on GenEval, especially for multi-object reasoning, highlighting the practicality of coupling understanding with generation.

Table 5 reports the results. We observe that jointly training for multimodal understanding consistently improves T2I generation quality. The gains are especially pronounced for the *Two Objects* category in GenEval, which requires resolving relationships across multiple entities—an ability naturally strengthened by the auxiliary understanding objective. These findings suggest that multimodal understanding and multimodal generation are mutually beneficial: enhancing the semantic reasoning capability of the MLLM leads to improved visual generation fidelity. This synergy points toward a promising direction for future unified architectures that treat understanding and generation as tightly coupled objectives rather than isolated tasks.

5. Limitation and Future Work

Limitation. A key limitation of our current design lies in its fixed output resolution of 512×512. While this choice simplifies training across heterogeneous tasks, it prevents UniGen-AR from flexibly adapting to inputs of arbitrary size. Adopting the dynamic-resolution strategies used in Stable Diffusion [53] and Infinity [21], *e.g.*, multiple groups of resolution choices and spatial padding, represents a practical next step toward broad deployment.

We additionally present typical **failure modes** in the supplementary material.

We have three **future research directions** following the current work: First, extending UniGen-AR to additionally handle editing tasks remains an important avenue, especially those requiring fine-grained, spatially localized modifications. Second, inspired by recent advances in MLLMs [39, 40] and diffusion transformers [5, 14], we aim to move from cross-attention conditioning toward a unified self-attention architecture, which we believe offers stronger coupling between modalities and improved controllability. Finally, our preliminary findings suggest that joint training of the MLLM and VAR decoder benefits multimodal alignment; we therefore see fully unified modeling – capable of both multimodal understanding (text-out) and generation (image-out) – as an exciting long-term goal.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 7

[2] Anurag Bagchi, Zhipeng Bao, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Refereverything: Towards segmenting everything we can speak of in videos. In *ICCV*, 2025. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 5, 7

[4] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, 2024. 3

[5] 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*, 2025. 2, 3, 8

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023. 6

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[8] Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. *arXiv preprint arXiv:2501.15420*, 2025. 3

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024. 5, 7

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2

[11] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *ECCV*, 2024. 6, 7

[12] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 8

[14] 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. 6, 8

[15] Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing. *arXiv preprint arXiv:2503.12652*, 2025. 3

[16] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, 2017. 6

[17] Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390*, 2023. 7

[18] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 3

[19] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023. 5, 8

[20] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2

[21] 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 *CVPR*, 2025. 2, 3, 4, 5, 6, 8

[22] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024. 3

[23] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. In *ICCV*, 2025. 5

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. In *NeurIPS*, 2017. 8

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[26] Zhipeng Huang, Shaobin Zhuang, Canmiao Fu, Binxin Yang, Ying Zhang, Chong Sun, Zhizheng Zhang, Yali Wang, Chen Li, and Zheng-Jun Zha. Wegen: A unified model for interactive multimodal generation as we chat. In *CVPR*, 2025. 3

[27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[28] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023. 3

[29] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024. 3

[30] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 7

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[32] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023. 3

[33] Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for few-shot image manipulation. In *CVPR*, 2025. 3

[34] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, 2022. 7

[35] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024. 3

[36] Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, Junhao Zhuang, Ying Shan, Yuexian Zou, and Qiang Xu. Brushedit: All-in-one image inpainting and editing. *arXiv preprint arXiv:2412.10316*, 2024. 3

[37] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. In *ICCV*, 2025. 2, 3, 5

[38] Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. *arXiv preprint arXiv:2503.10406*, 2025. 3

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 8

[40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 8

[41] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 2, 3

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[43] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3, 7

[44] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, 2024. 3

[45] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024. 3

[46] Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. *arXiv preprint arXiv:2508.15772*, 2025. 2, 3, 5

[47] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 5

[48] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 3

[49] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 3

[50] 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. *arXiv preprint arXiv:2307.01952*, 2023. 6

[51] Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution. *arXiv preprint arXiv:2501.18993*, 2025. 3

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 6

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 6, 8

[54] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[55] Chenze Shao, Fandong Meng, and Jie Zhou. Continuous visual autoregressive generation via score maximization. *arXiv preprint arXiv:2505.07812*, 2025. 3

[56] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[57] 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. 6

[58] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 3

[59] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun

Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024.

[60] Zeyi Sun, Ziyang Chu, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. X-prompt: Towards universal in-context image generation in auto-regressive vision language foundation models. *arXiv preprint arXiv:2412.01824*, 2024. 3, 6, 7

[61] Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv preprint arXiv:2503.21193*, 2025. 3

[62] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *CVPR*, 2024. 3

[63] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3, 6

[64] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 2024. 2, 3, 8

[65] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 3

[66] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024. 3, 4, 8

[67] Jiahuan Wang, Yuxin Chen, Jun Yu, Guangming Lu, and Wenjie Pei. Editinfinity: Image editing with binary-quantized generative models. *arXiv preprint arXiv:2510.20217*, 2025. 3

[68] Jinhong Wang, Jian Liu, Dongqi Tang, Weiqiang Wang, Wentong Li, Danny Chen, Jintai Chen, and Jian Wu. Scalable autoregressive monocular depth estimation. In *CVPR*, 2025. 3

[69] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3, 6

[70] Yufei Wang, Lanqing Guo, Zhihao Li, Jiaxing Huang, Pichao Wang, Bihan Wen, and Jian Wang. Training-free text-guided image editing with visual autoregressive model. *arXiv preprint arXiv:2503.23897*, 2025. 3

[71] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 5

[72] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *ICLR*, 2024. 5

[73] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, 2025. 3

[74] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2, 3, 5

[75] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 3

[76] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 3

[77] 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 *CVPR*, 2025. 3, 5, 7

[78] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 7

[79] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022. 7

[80] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 5

[81] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025. 3

[82] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *NeurIPS*, 2023. 3

[83] Shuhong Zheng, Zhipeng Bao, Ruoyu Zhao, Martial Hebert, and Yu-Xiong Wang. Diff-2-in-1: Bridging generation and dense perception with diffusion models. In *ICLR*, 2025. 2

[84] Hanshen Zhu, Zhen Zhu, Kaile Zhang, Yiming Gong, Yuliang Liu, and Xiang Bai. Training-free geometric image editing on diffusion models. In *ICCV*, 2025. 2

[85] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*, 2025. 3