Layout-Conditioned Autoregressive Text-to-Image Generation via Structured Masking

Zirui Zheng[1,2], Takashi Isobe[1✉], Tong Shen[1], Xu Jia[2✉], Jianbin Zhao[2], Xiaomin Li[1,2], Mengmeng Ge[1], Baolu Li[2], Qinghe Wang[2], Haiwen Diao[3], Dong Li[1], Dong Zhou[1], Yunzhi Zhuge[2], Huchuan Lu[2], and Emad Barsoum[1]

[1] Advanced Micro Devices Inc.
[2] Dalian University of Technology
[3] S-Lab, Nanyang Technological University

Fig. 1: Example results of layout-conditioned image generation produced by SMARLI. We show that SMARLI achieves fine-grained controllable generation with accurate spatial control and precise rendering of complex attributes, including color, shape, and texture. Global text prompts are omitted for brevity.

Abstract. Although autoregressive (AR) models have demonstrated remarkable success in image generation, extending these models to layout-conditioned generation remains challenging due to the sparse nature of layout conditions and the risk of feature entanglement. We present **S**tructured **M**asking for **AR**-based **L**ayout-to-**I**mage (SMARLI), a novel framework that effectively integrates spatial layout constraints into the AR generation process. To equip AR models with layout control, a structured masking strategy is applied to the attention computation to govern the interaction among the global prompt, layout, and image tokens. This design prevents the misassociation of different regions with their corresponding descriptions while enabling the sufficient injection of layout constraints into the generation process. To alleviate the exposure bias of AR models and further enhance generation quality and layout accuracy,

✉ Corresponding authors.

we incorporate a Group Relative Policy Optimization (GRPO) post-training scheme. We adapt it to the next-set-based paradigm and introduce a specifically designed layout reward, which is coordinated with an image quality reward to guide policy optimization in a balanced manner. Experimental results demonstrate that SMARLI seamlessly integrates layout tokens with text and image tokens without compromising generation quality, and the proposed masking strategy and post-training scheme can also be transferred to standard next-token-based AR models. The proposed framework achieves superior layout control while maintaining the structural simplicity and generation efficiency of AR models.

Keywords: Autoregressive Model · Layout-to-Image Generation · Post-training

1 Introduction

The success of AR modeling in Large Language Models [47,60] has inspired many methods that extend AR principles to image generation tasks [38,42,45,46,50,52, 55,56,63], showcasing the effectiveness of AR-based generative frameworks in the vision domain. These approaches reformulate image generation as a next-token, next-set, or next-scale prediction task, demonstrating remarkable potential and emerging as strong competitors to diffusion models. In this work, we investigate the potential of AR models in Layout-to-Image (L2I) generation.

The L2I generation conditions on the spatial locations and semantic descriptions of entities; it has been extensively explored in diffusion models based on different architectures, including U-Net [10, 12, 15, 30, 32, 51, 61, 68] and Multimodal Diffusion Transformers [24, 54, 65, 66]. Most existing methods for layout control in diffusion models integrate tokenized layout conditions by introducing additional blocks or even branches. However, such practices typically introduce substantial additional parameters and computational overhead. Furthermore, the introduction of additional blocks compromises the simplicity of the AR transformer architecture and conflicts with the emerging perspective of AR models as a unified generative framework. To avoid the structural modifications of the AR transformer, some prior studies [27,31,36,58,62] on AR-based conditional image generation adopt a feature-space injection paradigm inspired by ControlNet [67]. These methods incorporate encoded feature maps of visual condition by either adding them to image token embeddings or integrating them through joint decoding. However, they primarily rely on spatially aligned visual conditions that provide dense visual cues like edge maps or depth maps rather than layout. In contrast, layout conditions are inherently sparse and offer limited visual information, which may become further ambiguous when bounding boxes overlap. Moreover, jointly encoding bounding box coordinates and textual descriptions remains challenging, as it requires a dedicated condition encoder capable of modeling both geometric structure and semantic content.

To avoid specialized condition encoders that convert layouts into feature maps, an alternative paradigm concatenates conditional inputs with the text

prompt to form a unified sequence, enabling conditional image generation or image editing within AR models [9, 37, 56]. Despite this advancement, the exploration of this paradigm in L2I generation remains relatively limited. PlanGen [18] emerges as an early attempt that adopts this input sequence concatenation approach. By placing the tokenized layout conditions directly into the input sequence, it avoids complex architectural modifications. However, this approach faces two main challenges: (1) **semantic interference**, standard AR modeling fails to capture the structural characteristics of layout conditions. With a vanilla causal attention mask applied to the entire input sequence, layout tokens from different regions may interfere across regions, and image tokens may attend to irrelevant layout tokens. This results in feature entanglement and degraded layout precision, particularly in complex scenes. (2) **Limited image quality**, the visual fidelity of PlanGen remains inferior to that of state-of-the-art diffusion-based L2I models. The model tends to generate noticeable artifacts when synthesizing complex scenes. This limitation restricts its practical applicability in layout-conditioned image synthesis.

In this work, we propose SMARLI, an AR-based L2I framework that integrates a structured masking strategy to mitigate **semantic interference**, along with a layout-aware GRPO-based post-training scheme to improve **image quality** and layout controllability. Specifically, we first encode the global text prompt, layout condition, and image into a unified token sequence and feed it into the AR transformer. Then, to mitigate the aforementioned **interference**, we introduce the structured masking strategy, which explicitly regulates the attention computation across different token types: (1) Each object's layout tokens attend to the global text prompt tokens for broader contextual cues and to their local layout context, while remaining isolated from the layout tokens of other objects to prevent cross-object interference. (2) Image tokens receive guidance from both global text prompt tokens and region-specific layout tokens, ensuring the generation process is controlled by appropriate contexts while avoiding attribute confusion as illustrated in Fig. 1. This introduces inductive priors that align with the sequential organization of prompt, layout, and image tokens, facilitating efficient training without requiring architectural modifications to the base AR models. Finally, to further improve the **visual fidelity** and layout controllability, we propose a GRPO-based post-training scheme [44], and we adapt it to next-set-based AR models. A human preference score is introduced as a reward to enhance visual fidelity, and a comprehensive layout-aware reward is designed to improve layout controllability. Moreover, GRPO-based post-training mitigates exposure bias in AR models caused by the train–test discrepancy, as it trains on model-generated rollouts.

Extensive experiments demonstrate the effectiveness of the proposed framework in AR-based L2I generation. Furthermore, empirical results indicate that the proposed approach is also applicable to standard next-token-based AR models. It performs favorably against state-of-the-art diffusion-based counterparts, showing further potential of AR models.

Our main contributions are summarized as follows. (1) We introduce SMARLI, a layout-conditioned AR-based text-to-image (T2I) framework, which incorporates a structured masking strategy to inject layout conditions, mitigating region-description misalignment. (2) We adopt GRPO-based post-training and propose a comprehensive layout reward that, together with an image-quality reward, notably enhances both image fidelity and layout controllability. (3) Extensive experiments validate the effectiveness and generality of SMARLI across multiple L2I generation benchmarks.

2 Related Work

2.1 Autoregressive Image Generation

AR models have recently achieved image generation performance comparable to diffusion models. Typically, they employ a discrete VQ-VAE tokenizer [14] to convert images into sequences of tokens, which are then modeled using next-token prediction in raster-scan order [8, 45, 50]. To enhance both efficiency and output quality, recent approaches have reformulated the generation process. Masked generative models [4, 5, 13, 26, 52, 55] leverage bidirectional context to predict sets of masked tokens in parallel, while still maintaining AR training. VAR [46] introduces residual vector quantization to enable next-scale prediction, further improving image quality. In this paper, we present SMARLI, an AR-based image generation framework that endows AR models with explicit layout control. The proposed framework integrates global text prompts and spatial layout conditions into the autoregressive generation process, enabling precise spatial control without modifying the AR Transformer architecture.

2.2 Layout-conditioned T2I Generation

For layout control, existing diffusion-based T2I models often achieve effective layout control [7, 10, 12, 16, 17, 30, 51, 61, 65] by introducing additional attention between the layout and image or directly manipulating the attention map and image latent. For instance, GLIGEN [30] incorporates layout conditions into newly introduced trainable layers through a gated mechanism. InstanceDiffusion [51] extends this paradigm to instance-level control, supporting flexible layout specifications. SiamLayout [66] employs a separate branch of the network to process the layout in MMDiT. RPG [61] independently generates each region and composes them in the image latent space based on a pre-defined coarse layout. InstanceAssemble [54] proposed Assemble-MMDiT to inject the layout conditions into image tokens after global prompt injection. PlanGen [18] proposed a unified model for text-to-layout generation and layout-conditioned image generation. In this paper, based on the AR T2I model, we propose a structured masking strategy to realize effective layout control under the unified input sequence paradigm without introducing computationally expensive components.

2.3 Post-training in Visual Content Generation

The success of reward-based optimization in content generation [11, 28, 39, 57] has inspired extensive research on feedback-driven learning. Early works applied RL and preference optimization to diffusion models, including DDPO [3] and Diffusion-DPO [48]. Recent studies further generalized reward-based optimization across architectures, such as FlowGRPO [33], DanceGRPO [59], and ReNeg [29]. In next-token-based AR generation, SimpleAR [49] and T2I-R1 [22] employed GRPO post-training to improve text-image alignment.

However, GRPO has not been explored for AR-based layout-to-image (L2I) generation. To bridge this gap, we propose a GRPO-based post-training scheme with a specialized layout reward, complemented by an image quality reward, enabling more precise layout control while improving generation quality.

3 Method

We build our framework upon Show-o [55], which follows the next-set prediction paradigm. As shown in Fig. 2, our framework consists of three main components: (1) **Tokenization**: The global text prompt, layout, and image are tokenized separately, concatenated into a single sequence, and fed into the AR transformer. (2) **Layout control with structured masking strategy**: To equip the AR model with more effective layout control, a specially designed structured masking strategy is applied to attention computation. (3) **Layout GRPO**: To alleviate exposure bias and further improve the generation quality and layout accuracy, a GRPO-based post-training stage is adapted to the next-set-based AR model with an additional layout reward.

3.1 Preliminary

Next-set-based AR models Traditional AR models generate images token-by-token in raster order, leading to inefficiency for high-resolution images due to long sequences. Next-set-based AR models address this by predicting a set of image tokens at each timestep, reformulating image generation as masked token prediction. Specifically, during inference, all image tokens are initially masked, and at each time step, the model predicts a subset of tokens in parallel. Masked tokens are progressively revealed following a scheduling strategy that reduces the masking ratio from 1.0 to 0.

GRPO We adopt GRPO algorithm [44] for post-training, which extends PPO (Proximal Policy Optimization [43]) by removing the value function and estimating advantages in a group-relative manner. The objective is defined as:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{c\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left(r_{i,t}A_{i,t},\, \hat{r}_{i,t}A_{i,t}\right) - \beta\, \mathrm{D}_{\mathrm{KL}}\right\}\right] \tag{1}$$

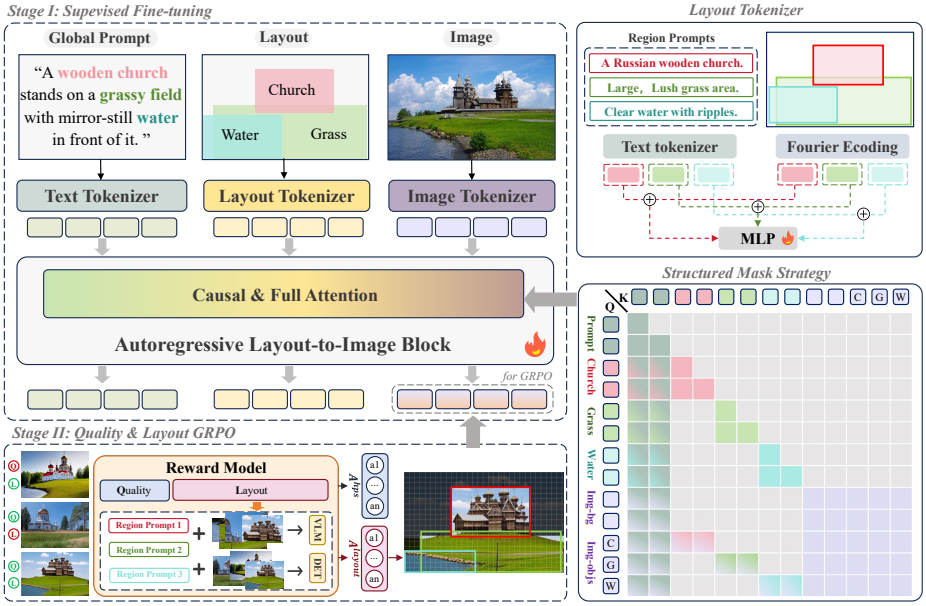


Fig. 2: An overview of the proposed framework. In stage 1, the global prompt and image are tokenized using Show-o's original setup. Layout tokens are derived from the layout tokenizer based on the bounding boxes and region prompts. A unified input sequence, formed by the three types of tokens, is fed into the transformer equipped with a dedicated structured masking strategy to enable more effective layout control. After SFT, we further introduce a GRPO-based post-training stage to improve the layout controllability and image quality simultaneously with a dedicated layout reward and image quality reward.

GRPO-based post-training involves a trainable policy model π_θ, a behavior policy model $\pi_{\theta_{old}}$, and a frozen reference model π_{ref}, all initialized from a checkpoint obtained through prior supervised fine-tuning (SFT). Given condition c from training data $\mathcal{D}$, the behavior policy model $\pi_{\theta_{old}}$ samples a group of outputs $o_1, o_2, ..., o_G$, and then optimizes the objective in Eq. (1) to update π_θ. The $r_{i,t} = \frac{\pi_\theta(o_{i,t}|c,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|c,o_{i,<t})}$ denotes the importance sampling ratio between the trainable policy model π_θ and behavior policy model $\pi_{\theta_{\text{old}}}$, while $\hat{r}_{i,t} = \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)$ applies a clipping threshold ϵ to stabilize training. $A_{i,t}$ represents the advantage of the t-th token of o_i obtained by group-wise normalized rewards, since action for the traditional AR model is every single token. The KL divergence $\text{D}_{\text{KL}} = \text{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$, weighted by coefficient β, regularizes the updated policy towards a fixed reference policy π_{ref} to reduce overfitting.

3.2 Tokenization

For the prompt and image, we adopt Show-o's setup: the prompt is tokenized by Show-o's text tokenizer, and the image is tokenized into discrete tokens by

MAGVITv2 [64]. Layout condition consists of multiple objects, where each object is defined by a bounding box $\text{obj}_i^{\text{box}} = [x_0, y_0, x_1, y_1]$ and a textual description $\text{obj}_i^{\text{text}}$. The box specifies the normalized top-left and bottom-right coordinates, while the text provides the semantic description.

To integrate geometric and linguistic information, we adopt a unified layout encoding strategy that embeds spatial boxes and their textual descriptions together. Specifically, we directly use Show-o's text tokenizer E and codebook C to encode the region description of the layout; each bounding box is represented as a set of Fourier embeddings. Unlike diffusion models that concatenate text tokens with Fourier embeddings along the channel dimension, we aim to fully leverage the text representations learned during Show-o pretraining. To this end, we concatenate the text tokens and box tokens along the sequence dimension and process them with a zero-initialized MLP equipped with a residual connection, as illustrated in the Layout Tokenizer of Fig. 2.

$$x_i = \text{Concat}\left(C(E(obj_i^{\text{text}})),\ \text{Fourier}(obj_i^{\text{box}})\right) \tag{2}$$

$$obj_i^{\text{layout}} = x_i + \text{MLP}(x_i) \tag{3}$$

The layout tokens of object i can be expressed as Eqs. (2) and (3), where *Concat* denotes concatenation in the sequence dimension. Finally, the input to the AR transformer is a sequence composed of the global text prompt, layout, and image tokens, where the image tokens are partially replaced by [MASK] tokens for SFT training.

3.3 Structured Masking Strategy

As shown in Fig. 2, given a sequence of tokens for the prompt, layout, and image tokens in order, a specially designed structured masking strategy is applied to attention computation in transformer blocks as follows.

As for the global text prompt tokens, we adopt the standard causal masking strategy. As for the layout tokens, our attention masking strategy follows three design principles: (1) **Global Context Awareness**, each layout token is allowed to attend to the global text prompt tokens, enabling it to capture high-level contextual cues and better understand how a specific region relates to the entire image. (2) **Intra-object Coherence**, layout tokens that belong to the same object are connected under a standard causal mask, allowing them to accumulate fine-grained spatial and semantic information. (3) **Inter-object Isolation**, for layout tokens belonging to different objects, the masks are designed to isolate them from each other to avoid information interference and preserve object-specific representations.

For image tokens, they attend to all other image tokens and the global prompt tokens, but only to the subset of layout tokens they are associated with, following the principle of **Region-wise Separation.** If an image token lies within multiple overlapping object boxes, it attends to the associated layout tokens to capture composite layout semantics. In such cases, the global text prompt provides relational context between the overlapping objects, helping the model

interpret their semantic relationship and resolve the resulting ambiguity. In this way, the generation process of image tokens is steered by global semantic guidance and associated layout conditions, resulting in precise spatial and semantic alignment while mitigating interference from irrelevant tokens.

Together, these strategies introduce a structured inductive prior that aligns with the sequential organization of prompt, layout, and image tokens, facilitating efficient training without introducing notable architectural modifications or significant parameter overhead.

3.4 Layout-GRPO

To further improve image quality and layout controllability, we adopt a GRPO-based post-training framework. However, adapting GRPO to next-set-based AR models for layout-conditioned image generation poses unique challenges. (1) How to define the *action* for next-set-based AR models in GRPO, (2) how to enhance image fidelity without compromising, and potentially improving, the layout controllability of the SFT model.

As introduced in Sec. 3.1, Show-o follows a *next-set prediction* paradigm. At each time step, the model attends to the entire image token sequence and retains only a subset of tokens with high confidence as predictions, while the remaining tokens are masked again, so the retained token set at each step should be regarded as part of the action. In standard AR models following the next-token-predication paradigm, the context at each time step is simply the generated token prefix, since all prefixes are nested subsequences of the same full sequence, a single forward pass with a causal attention mask can simultaneously construct all step-wise contexts, so the log-probabilities for all time steps can be computed in parallel within one forward evaluation. In contrast, Show-o uses the full-length input token sequence as context at every step, and the contexts may differ in the token values at the same spatial positions. Such step-dependent and full-length contexts cannot be reproduced by a fixed attention mask, making it impractical to compute all step-wise log-probabilities in a single forward pass. To address this issue, during rollouts with $\pi_{\theta_{\text{old}}}$, we explicitly record the retained token set together with their spatial positions at each time step as a composite action, and treat the current input token sequence as the corresponding state. The π_θ and $\pi_{\theta_{\text{ref}}}$ then replay the same state–action trajectory step by step, allowing us to compute $\pi_\theta(o_{i,t} \mid c, o_{i,<t})$ and $\pi_{\theta_{\text{ref}}}(o_{i,t} \mid c, o_{i,<t})$ under identical contexts.

To preserve layout controllability while improving the image quality, we propose a dedicated layout reward that integrates well with the image quality reward (HPS v2.1 score [53]), leading to improved overall performance. Specifically, for each bounding box in the layout, we evaluate the corresponding image region with its detailed textual description in a visual question answering (VQA) manner using a vision-language model [1] (VLM), obtaining a binary **VQA score** S_{VQA} that reflects semantic and attribute consistency. However, a positive VQA score is given as long as the object's main part appears in the specified region, regardless of precise spatial boundaries, and it does not penalize false positives when the object appears elsewhere. To address this, we further introduce the

mIoU score and the precision score to encourage more accurate locations and avoid false positives. Specifically, we employ Grounding DINO [34] to detect the generated image according to the given category labels, and compute the mIoU and precision by comparing the detected results with the layout conditions. The weighted sum of these score forms the layout reward R_{layout}, as shown in Eq. (4).

$$R_{\text{layout}} = \lambda_{\text{VQA}} \cdot S_{\text{VQA}} + \lambda_{\text{mIoU}} \cdot S_{\text{mIoU}} + \lambda_{\text{prec}} \cdot S_{\text{prec}}, \tag{4}$$

$$A_{i,t} = \omega^{\text{hps}} \cdot A_i^{\text{hps}} + \omega^{\text{layout}} \cdot A_i^{\text{layout}} \tag{5}$$

The overall advantage is defined in Eq. (5) as a weighted combination of the HPS advantage A_i^{hps} and the layout advantage A_i^{layout}. A larger ω^{layout} is further applied to image tokens located within the bounding boxes specified by the layout condition, thereby encouraging the model to focus on layout-aligned regions during policy optimization.

4 Experiments

4.1 Datasets and Evaluation Metrics

Data Preparation. For SFT training, we use the LayoutSAM [66] dataset. For GRPO-based post-training, to ensure reliable layout rewards, we filter out samples from LayoutSAM that contain semantically redundant instances, and we further remove samples whose Grounding-DINO [34] detection results fail to reach mIoU and precision above 0.9.

Benchmark and Evaluation Metric. We evaluate our method across various benchmarks, including the Layout-to-Image (L2I) datasets LayoutSAM-Eval [66] and OverLayBench [25], as well as the compositional Text-to-Image (T2I) benchmark T2I-CompBench [21]. For LayoutSAM-Eval, we assess layout control through spatial positioning and attribute alignment, specifically focusing on color, texture, and shape consistency. Image quality is measured by human preference scores (IR [57], Pick [23]), text-image alignment (CLIP [41]), and perceptual quality (FID [19]). Furthermore, 895 samples are selected from LayoutSAM-Eval using Grounding-DINO to evaluate mIoU and precision. OverLayBench addresses more demanding layout conditions characterized by multiple overlapping bounding boxes with semantically similar categories. On this benchmark, layout control is evaluated via mIoU, O-mIoU (mIoU calculated on overlapped regions), SRE (Success Rate of Entity), and SRR (Success Rate of generated spatial Relationships). Semantic consistency is measured at both image and region levels using CLIP_{G} and CLIP_{L}, respectively. Regarding T2I-CompBench, we use metrics that intersect with layout-related capabilities, including color, shape, texture, spatial reasoning, and numeracy.

4.2 Training Details

We adopt Show-o as the foundation model, with a training resolution of 512×512. For the SFT stage, we use the AdamW [35] optimizer with a fixed learning

rate of 2×10^{-5}, training for 30,000 iterations with a global batch size of 128. Trainable components include the entire AR transformer and the MLP in the layout tokenizer. The SFT stage took 3 days on 8 AMD MI300X GPUs.

In GRPO-based post-training, the number of groups is set to 4, the number of generation steps to 10, the batch size to 28, and the learning rate to 1×10^{-4}, with training conducted for 100 iterations. Each transformer block is fine-tuned using LoRA [20] with a rank of 256. For rollouts, the temperature is set to 1 and the classifier-free guidance scale to 5. Following DanceGRPO [59], to mitigate the training instability introduced by rollouts with classifier-free guidance, each policy update relies exclusively on newly generated rollouts, and β in Eq. (1) is set to 0. The weights ω^{hps} and ω^{layout} are set to 1. When image tokens are located within layout-specified bounding boxes, ω^{layout} is increased to 1.2 to reinforce layout control. For the layout reward in Eq. (4), λ_{VQA}, λ_{mIoU}, and λ_{prec} are set to 0.5, 0.5, and 0.1, respectively. The GRPO stage took 4 hours on 8 AMD MI250 GPUs. Compared with the SFT stage, the GRPO stage introduces only **minimal** additional training overhead, while yielding clear improvements in both image quality and layout controllability.

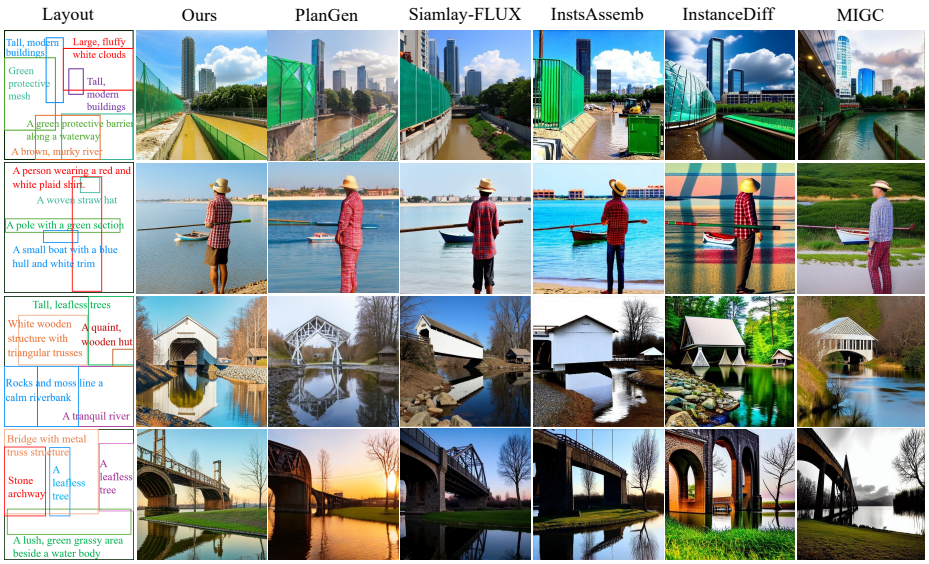


Fig. 3: Qualitative results on LayoutSAM-Eval. Under complex layouts, previous methods often fail to generate all specified objects or capture their fine-grained attributes, whereas SMARLI reliably adheres to the layout and generates all required objects.

4.3 Evaluation on Layout-to-Image Generation

We compare SMARLI with state-of-the-art L2I models on LayoutSAM-Eval and OverLayBench. For LayoutSAM-Eval, most results are adopted from CreatiLayout [66], while the scores of InstsAssemb and PlanGen are taken from their

original papers. For OverLayBench, we use the official benchmark results when available and reproduce the results of InstsAssemb and PlanGen using their released code. In addition, we validate SMARLI on another autoregressive image generation model, Janus-Pro-1B, which follows standard next-token prediction. It achieves performance comparable to SMARLI-Show-o. Detailed results and analysis are provided in the supplementary material.

Table 1: Quantitative results on LayoutSAM-Eval [66]. Spa, Col, Tex, Sha, and Pre denote Spatial, Color, Texture, Shape, and Precision, respectively. **Bold** and underline indicate the best and second-best results. * means it is trained on LayoutSAM.

Method	Params	Layout Control						Image Quality			
		Spa.↑	Col.↑	Tex.↑	Sha.↑	mIoU↑	Pre.↑	IR↑	Pick↑	CLIP↑	FID↓
GT Images	-	98.95	98.45	98.90	98.80	98.36	100.0	-	-	-	-
GLIGEN [30]	1.2B	77.53	49.41	55.29	52.72	69.62	70.99	-10.31	20.78	32.42	21.92
InstanceDiff [51]	1.4B	87.99	69.16	72.78	71.08	81.29	72.34	9.14	21.01	31.40	19.67
MIGC [68]	1.2B	85.66	66.97	71.24	69.06	70.99	74.10	-13.72	20.71	31.36	21.19
HiCo [10]	1.4B	87.04	69.19	72.36	71.10	81.20	71.74	12.36	21.70	32.18	22.61
SiamLay-SD3* [66]	3B	92.67	74.45	77.21	75.93	68.43	63.22	69.47	22.02	**34.01**	19.10
InstsAssemb* [54]	2B	94.97	77.53	80.72	80.11	73.33	74.19	70.14	21.75	33.08	20.24
SiamLay-Flux* [66]	22B	**95.67**	80.71	83.53	82.80	78.01	78.33	**80.48**	**22.16**	33.92	16.12
PlanGen* [18]	1.5B	92.21	82.69	86.53	85.36	78.04	81.15	39.58	21.43	31.96	**13.91**
SMARLI*	1.3B	95.55	**87.35**	**90.90**	**89.82**	**81.31**	**81.84**	74.81	21.81	33.14	18.16

Table 2: Quantitative results on OverLayBench-Complex.

Method	Params	mIoU ↑	O-mIoU ↑	SR_E ↑	SR_R ↑	$CLIP_G$ ↑	$CLIP_L$ ↑
GLIGEN [30]	1.2B	50.79	23.85	41.70	79.93	33.92	22.75
InstanceDiff [51]	1.4B	53.68	25.63	66.02	80.34	32.33	25.53
MIGC [68]	1.2B	40.04	13.26	47.80	74.48	31.93	24.20
HiCo [10]	1.4B	46.56	20.35	48.88	75.19	33.15	24.41
SiamLay-SD3 [66]	3B	44.24	18.05	52.10	79.98	36.55	24.76
SiamLay-FLUX [66]	22B	54.50	28.97	69.72	86.45	**36.72**	24.85
EliGen [65]	12B	52.53	26.19	74.03	84.09	36.18	25.92
InstsAssemb [54]	2B	49.12	23.26	58.55	83.22	35.29	24.32
PlanGen [18]	1.5B	51.21	23.53	67.79	85.36	35.22	19.17
SMARLI	1.3B	**56.42**	**29.59**	**80.52**	**87.81**	36.57	**26.48**

The quantitative results are presented in Tab. 1 and Tab. 2. On LayoutSAM-Eval, the SMARLI demonstrates strong layout control performance, excelling in spatial, color, texture, and shape control. It also achieves promising quality performance, with competitive results in these image quality metrics. On OverlayBench-Complex, SMARLI maintains more robust layout alignment under layouts with multiple overlapping bounding boxes. For completeness, results

on the Simple and Regular subsets are provided in the supplementary material, where SMARLI consistently outperforms other methods, further demonstrating its effectiveness across varying layout complexities.

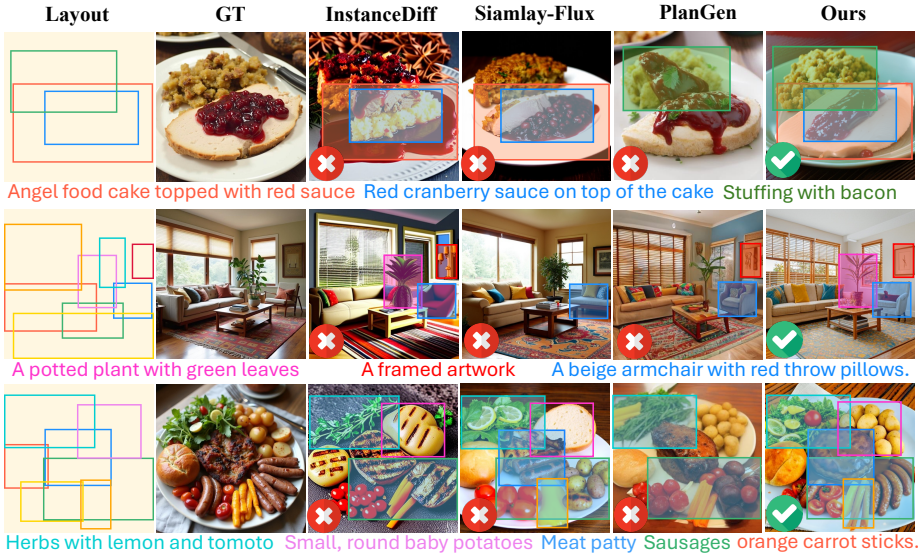


Fig. 4: Qualitative results on OverLayBench.

The qualitative results are shown in Fig. 3 and Fig. 4. Images generated by our method demonstrate superior layout control, exhibiting more accurate spatial positioning and faithful attribute rendering than existing approaches. These results highlight the effectiveness of the proposed masking strategy, which enhances the semantic representation of layout tokens and mitigates interference between tokens from different objects, together with the benefits introduced by the well-designed layout reward in GRPO-based post-training. The compelling quantitative and qualitative results demonstrate the potential of AR models as strong competitors to diffusion models.

4.4 Evaluation on Text-to-Image Generation

Following Creatilayout [66], we conduct experiments on T2I-CompBench to evaluate the benefit brought by the additional layout condition. Since the T2I-CompBench only provides the text prompts, we use GPT-5.2 to generate

Table 3: Quantitative results on T2I-CompBench.

Model	Spatial ↑	Color ↑	Shape ↑	Texture ↑	Numeracy ↑
SDXL [40]	21.3	58.8	46.9	53.0	49.9
PixArt-α [6]	20.6	66.9	49.3	64.8	50.6
DALLE-3 [2]	28.7	<u>77.9</u>	**62.1**	<u>70.4</u>	58.8
Show-o [55]	<u>39.6</u>	73.0	52.4	67.7	<u>62.3</u>
SMARLI	**49.7**	**79.6**	<u>59.2</u>	**72.3**	**68.4**

the corresponding layouts. As shown in Tab. 3, conditioned on the text prompt and layouts, SMARLI achieves a significant improvement in spatial adherence (from 39.6 to 49.7). Additionally, the benefits of the layout are also reflected in the consistent improvements in numeracy and attribute binding.

4.5 Ablation Study

In this section, we conduct ablation studies on the structure masking strategy and GRPO-based Post-training. Further ablation studies on the layout tokenizer and ω^{layout} can be found in the supplementary material.

Masking Strategy. We first conduct an ablation study on the effectiveness of the proposed masking strategy. The model without GRPO-based post-training (SMARLI-SFT) is used as the baseline, and the results are reported in Tab. 4. We examine two variant masking strategies: (1) layout tokens are isolated from global prompt tokens (w/o LP); (2) layout tokens are no longer treated as independent units per object but are instead processed as a whole (w/o Local Causal).

Table 4: Ablation study on different masking strategies. Compared to the baseline, a noticeable performance degradation in spatial positioning and attribute binding is observed when the LP or Local Causal is removed.

Mask Strategy	Spa.	Col.	Tex.	Sha.	mIoU	Pre.
SMARLI-SFT	**95.26**	**86.17**	**89.79**	**88.56**	80.08	**80.69**
w/o LP	95.17	79.83	84.35	82.74	**80.13**	80.55
w/o Local Causal	93.93	83.22	87.12	85.68	78.39	77.04

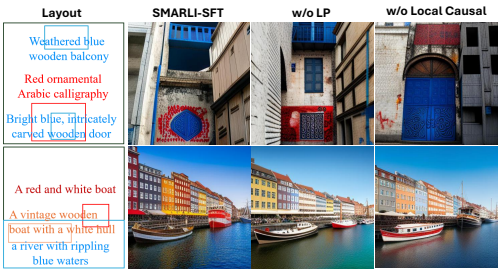


Fig. 5: Visualization of ablation study on masking strategies.

For the w/o LP, layout tokens are prohibited from attending to global prompt tokens. As a result, layout tokens cannot leverage the contextual semantic information provided by global prompt tokens, significantly undermining the model's ability to control object attributes. As illustrated in Fig. 5, it causes failures in attribute rendering, highlighting the critical role of Global Context Awareness.

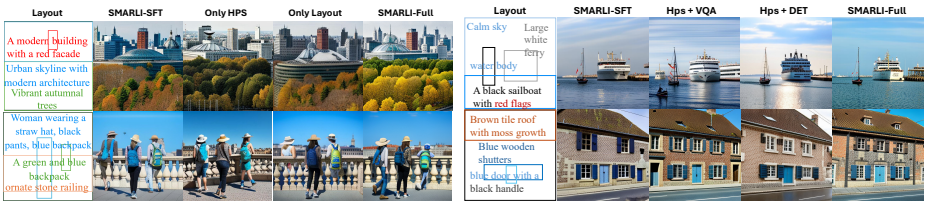
For w/o Local Causal, layout tokens attend to both global prompt tokens and preceding layout tokens, while image tokens are not restricted to their corresponding regions. This indiscriminate mixing of attribute and spatial information across objects degrades performance, especially in spatial metrics. As shown in Fig. 5, it causes misaligned object spatial positioning and attribute confusion, underscoring the importance of Intra-object Coherence, Inter-object Isolation, and Region-wise Separation.

GRPO-based Post-training. We conduct ablation studies starting from an SFT baseline to evaluate GRPO-based post-training. As shown in Tab. 5 and

Table 5: Ablation study on GRPO-based post-training.

HPS	VQA	DET	Layout Control						Image Quality	
			Spatial	Color	Texture	Shape	mIoU	Pre	IR	Pick
✗	✗	✗	95.26	86.17	89.79	88.56	80.08	80.69	60.88	21.45
✓	✗	✗	93.82	83.96	87.72	86.78	77.79	77.89	**76.62**	**21.96**
✗	✓	✓	95.41	87.04	90.15	89.32	80.76	80.81	58.34	21.28
✓	✓	✗	95.48	87.30	90.01	88.97	77.73	74.50	73.36	21.69
✓	✗	✓	95.13	85.61	89.45	88.29	80.70	80.88	73.05	21.53
✓	✓	✓	**95.55**	**87.35**	**90.90**	**89.82**	**81.31**	**81.84**	74.81	21.81

Fig. 6a, using HPS alone ("Only HPS") improves image quality but reduces layout consistency, particularly in color accuracy. Using layout reward only ("Only Layout") enhances layout alignment but degrades the image quality. Combining both rewards ("SMARLI-Full") achieves improved visual quality and layout adherence, with HPS further reinforcing global prompt-image consistency.



(a) Ablation study on hps and layout rewards. **(b)** Detailed ablation study on layout reward.

Fig. 6: Visualization of ablation study on rewards.

Moreover, we ablate the layout reward design. We divide the layout reward into the VQA score and the detection (DET) score (mIoU score and precision score). As shown in Tab. 5 and Fig. 6b, together with hps score, using only the VQA score ("Hps+VQA") improves spatial positioning and attribute alignment but neglects low-IoU and false-positive issues. Conversely, using only the detection score ("Hps+DET") effectively reduces low-IoU and false positives, yet it compromises semantic consistency with the detailed textual description. The combination of both rewards leads to more balanced and superior performance.

5 Conclusion

In this paper, we have presented SMARLI, a novel framework that effectively integrates layout constraints into AR-based T2I generation. Our unified input

design and structured masking strategy enable precise spatial control and coherent attribute rendering, while GRPO-based post-training with image-quality and layout rewards further improves generation fidelity and layout alignment. Experimental results demonstrate that SMARLI achieves superior layout control while maintaining the simplicity of AR models. This work provides valuable insights for incorporating layout conditions into AR models, advancing the field of AR-based controllable image generation.

6 Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 62441231, 62472065, U23B2010, Talent Fund of Liaoning Province under Grant No. XLYC2503161, and the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No. GML-KF-26-08.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
3. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
4. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
5. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
6. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
7. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5343–5353 (2024)
8. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)

9. Chen, Y., Ma, Z., Jia, G., Jiang, C., Li, J., Zhou, B.: Context-aware autoregressive models for multi-conditional image generation. arXiv preprint arXiv:2505.12274 (2025)
10. Cheng, B., Ma, Y., Wu, L., Liu, S., Ma, A., Wu, X., Leng, D., Yin, Y.: Hico: Hierarchical controllable diffusion model for layout-to-image generation. arXiv preprint arXiv:2410.14324 (2024)
11. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. arXiv preprint arXiv:2309.17400 (2023)
12. Dahary, O., Patashnik, O., Aberman, K., Cohen-Or, D.: Be yourself: Bounded attention for multi-subject text-to-image generation. In: European Conference on Computer Vision. pp. 432–448. Springer (2024)
13. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169 (2024)
14. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
15. Feng, Y., Gong, B., Chen, D., Shen, Y., Liu, Y., Zhou, J.: Ranni: Taming text-to-image diffusion for accurate instruction following. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4744–4753 (2024)
16. Ge, M., Isobe, T., Jia, X., Sun, Y., Yang, Z., Wang, W., Zhou, D., Li, D., Lu, H., Barsoum, E.: Ego-inbetween: Generating object state transitions in ego-centric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 43145–43154 (2026)
17. Ge, M., Jia, X., Isobe, T., Li, X., Wang, Q., Mu, J., Zhou, D., Wang, L., Lu, H., Tian, L., et al.: Customizing text-to-image generation with inverted interaction. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10901–10909 (2024)
18. He, R., Cheng, B., Ma, Y., Jia, Q., Liu, S., Ma, A., Wu, X., Wu, L., Leng, D., Yin, Y.: Plangen: Towards unified layout planning and image generation in auto-regressive vision language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18143–18154 (2025)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
21. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
22. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025)
23. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems **36**, 36652–36663 (2023)
24. Lee, Y., Yoon, T., Sung, M.: Groundit: Grounding diffusion transformers via noisy patch transplantation. Advances in Neural Information Processing Systems **37**, 58610–58636 (2024)

25. Li, B., Wang, C.Y., Xu, H., Zhang, X., Armand, E.J., Srivastava, D., Shan, X., Chen, Z., Xie, J., Tu, Z.: Overlaybench: A benchmark for layout-to-image generation with dense overlaps. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
26. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
27. Li, X., Qiu, K., Chen, H., Kuen, J., Lin, Z., Singh, R., Raj, B.: Controlvar: Exploring controllable visual autoregressive modeling. arXiv preprint arXiv:2406.09750 (2024)
28. Li, X., Liang, Q., Li, Y., Zhang, Y., Li, C., Lyu, J., Lu, H., Jia, X.: Portraitgen: Exemplar-driven grpo with dual-reward guidance for photorealistic portrait generation (2026), https://arxiv.org/abs/2606.26930
29. Li, X., Liu, Y., Isobe, T., Jia, X., Cui, Q., Zhou, D., Li, D., He, Y., Lu, H., Wang, Z., et al.: Reneg: Learning negative embedding with reward guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23636–23645 (2025)
30. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
31. Li, Z., Cheng, T., Chen, S., Sun, P., Shen, H., Ran, L., Chen, X., Liu, W., Wang, X.: Controlar: Controllable image generation with autoregressive models. arXiv preprint arXiv:2410.02705 (2024)
32. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023)
33. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
34. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., 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)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
36. Mao, Q., Cai, Q., Li, Y., Pan, Y., Cheng, M., Yao, T., Liu, Q., Mei, T.: Visual autoregressive modeling for instruction-guided image editing. arXiv preprint arXiv:2508.15772 (2025)
37. Mu, J., Vasconcelos, N., Wang, X.: Editar: Unified conditional generation with autoregressive models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7899–7909 (2025)
38. Mumuni, A., Mumuni, F., Gerrar, N.K.: A survey of synthetic data augmentation methods in machine vision. Machine Intelligence Research **21**(5), 831–869 (2024)
39. Peng, S., Wang, W., Tian, Z., Yang, S., Wu, X., Xu, H., Zhang, C., Isobe, T., Hu, B., Zhang, M.: Omni-dpo: A dual-perspective paradigm for dynamic preference learning of llms. arXiv preprint arXiv:2506.10054 (2025)
40. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
42. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
43. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
44. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
45. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
46. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
47. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
48. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
49. Wang, J., Tian, Z., Wang, X., Zhang, X., Huang, W., Wu, Z., Jiang, Y.G.: Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. arXiv preprint arXiv:2504.11455 (2025)
50. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
51. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6232–6242 (2024)
52. Wu, S., Zhang, W., Xu, L., Jin, S., Wu, Z., Tao, Q., Liu, W., Li, W., Loy, C.C.: Harmonizing visual representations for unified multimodal understanding and generation. arXiv preprint arXiv:2503.21979 (2025)
53. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
54. Xiang, Q., Sun, S., Li, B., Song, D., Li, H., Chen, N., Tang, X., Hu, Y., Zhang, J.: Instanceassemble: Layout-aware image generation via instance assembling attention. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
55. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
56. Xin, Y., Yan, J., Qin, Q., Li, Z., Liu, D., Li, S., Huang, V.S.J., Zhou, Y., Zhang, R., Zhuo, L., et al.: Lumina-mgpt 2.0: Stand-alone autoregressive image modeling. arXiv preprint arXiv:2507.17801 (2025)

57. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
58. Xu, R., Jin, D., Bai, Y., Lan, R., Duan, X., Sun, L., Chu, X.: Scalar: Scale-wise controllable visual autoregressive learning. arXiv preprint arXiv:2507.19946 (2025)
59. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
60. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
61. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: Forty-first International Conference on Machine Learning (2024)
62. Yao, Z., Li, J., Zhou, Y., Liu, Y., Jiang, X., Wang, C., Zheng, F., Zou, Y., Li, L.: Car: Controllable autoregressive modeling for visual generation. arXiv preprint arXiv:2410.04671 (2024)
63. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved vqgan. arXiv preprint arXiv:2110.04627 (2021)
64. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023)
65. Zhang, H., Duan, Z., Wang, X., Chen, Y., Zhang, Y.: Eligen: Entity-level controlled image generation with regional attention. arXiv preprint arXiv:2501.01097 (2025)
66. Zhang, H., Hong, D., Wang, Y., Shao, J., Wu, X., Wu, Z., Jiang, Y.G.: Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. arXiv preprint arXiv:2412.03859 (2024)
67. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
68. Zhou, D., Li, Y., Ma, F., Zhang, X., Yang, Y.: Migc: Multi-instance generation controller for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6818–6828 (2024)