# SCALAR: Scale-wise Controllable Visual Autoregressive Learning

**Ryan Xu*, Dongyang Jin*, Yancheng Bai†, Rui Lan, Xu Duan, Lei Sun‡, and Xiangxiang Chu**

Amap, Alibaba Group
ryansxu.00@gmail.com, {jindongyang.j, lr264907, xuxu.dx}@alibaba-inc.com,
{yancheng.byc, ally.sl, chuxiangxiang.cxx}@alibaba-inc.com

## Abstract

Controllable image synthesis, which enables fine-grained control over generated outputs, has emerged as a key focus in visual generative modeling. However, controllable generation remains challenging for Visual Autoregressive (VAR) models due to their hierarchical, next-scale prediction style. Existing VAR-based methods often suffer from inefficient control encoding and disruptive injection mechanisms that compromise both fidelity and efficiency. In this work, we present SCALAR, a controllable generation method based on VAR, incorporating a novel Scale-wise Conditional Decoding mechanism. SCALAR leverages a pretrained image encoder to extract semantic control signal encodings, which are projected into scale-specific representations and injected into the corresponding layers of the VAR backbone. This design provides persistent and structurally aligned guidance throughout the generation process. Building on SCALAR, we develop SCALAR-Uni, a unified extension that aligns multiple control modalities into a shared latent space, supporting flexible multi-conditional guidance in a single model. Extensive experiments show that SCALAR achieves superior generation quality and control precision across various tasks.

**Code** — https://github.com/AMAP-ML/SCALAR

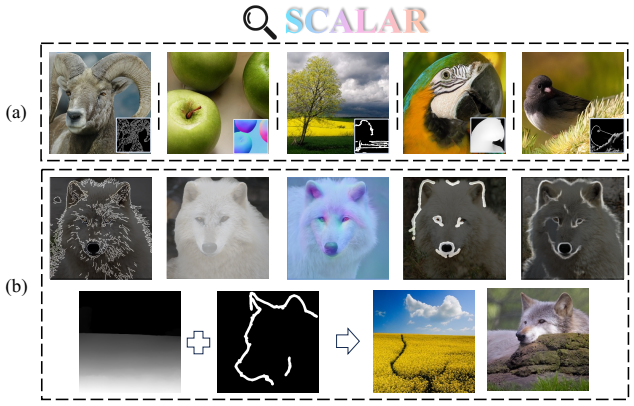


Figure 1: (a) **SCALAR**, a novel controllable VAR method with superior generation quality and control (top row). (b) **SCALAR-Uni** further extends it by supporting multi-condition control within a unified model (bottom row). Dashed lines denote different model weights.

## Introduction

Controllable image synthesis is a pivotal domain in visual generation, enabling the precise and nuanced creation of visual content according to specific user guidance. Recent advances in this field are currently dominated by two primary paradigms: Diffusion Models (Ho, Jain, and Abbeel 2020; Song et al. 2020) and Autoregressive Models (AR) (Sun et al. 2024; Tian et al. 2024). While diffusion-based methods (Zhang, Rao, and Agrawala 2023a; Qin et al. 2023) have achieved widespread success, the iterative denoising process is not inherently compatible with the sequential, token-based architecture of LLMs, which hinders the development of truly unified multimodal modeling (Zhao et al. 2023).

Visual Autoregressive (VAR) models (Tian et al. 2024), a leading approach within the autoregressive paradigm, offer a compelling path forward. By framing image synthesis as a next-scale prediction task, VAR-based models (Han et al. 2025; Tang et al. 2024; Ma et al. 2024; Zhuang et al. 2025) align naturally with LLM architectures and have demonstrated better inference efficiency and generative quality than both state-of-the-art diffusion models (Podell et al. 2023) and raster-scan-based autoregressive models (Sun et al. 2024). Nevertheless, the capacity of VAR models for fine-grained control remains a significant and underexplored challenge. This challenge primarily stems from the unique hierarchical, scale-wise generation process, which presents a distinct paradigm from existing approaches. In diffusion models, control signals are applied globally to the denoising network at each step, whereas in traditional raster-scan AR models (Sun et al. 2024; Li et al. 2024a), they are injected before each spatial token prediction.

Existing works on controllable Visual Autoregressive learning, notably CAR (Yao et al. 2024) and ControlVAR (Li et al. 2024c), deliver suboptimal performance. We attribute this bottleneck to two fundamental design flaws. First, they utilize complex and disruptive injection mechanisms; architectures with parallel branches (Yao et al. 2024) similar to

*These authors contributed equally.
†Corresponding author and project leader
‡Corresponding author

ControlNet or joint control and image modeling schemes (Li et al. 2024c) not only incur substantial computational overhead but also implicitly disrupt the powerful generative capabilities of the pretrained backbone. Second, they often utilize lightweight convolutional networks or VQ-VAEs as control encoders, which have been proven to have a limited capacity in capturing rich spatial-semantic features (Zhou et al. 2024). Consequently, their performance in both generation quality and control consistency trails behind even raster-scan AR models (Li et al. 2024d), highlighting a critical need for a control mechanism that is both efficient and fundamentally aligned with the native structure of VAR.

To this end, we present SCALAR, an effective controllable generation method based on VAR. Aligning with the more efficient paradigm (Li et al. 2024d), SCALAR injects control during the autoregressive decoding phase. Our approach is centered on a novel and simple **Scale-wise Conditional Decoding** mechanism. To be specific, we first employ a pretrained vision foundation model (Oquab et al. 2023) to extract powerful, scale-agnostic control features that contain rich semantic information. These general features are then processed by scale-wise lightweight projection blocks—which maintain independent weights for each scale—to produce specialized **Control Signal Encodings**. Ultimately, SCALAR injects these tailored encodings directly into the hidden states of the VAR backbone's layers, providing scale-wise persistent guidance throughout the next-scale prediction process. Building on our design, we extend SCALAR to a unified control version, SCALAR-uni. This extension incorporates a **Unified Control Alignment** process that projects features from diverse control modalities into a common latent space, enabling seamless guidance of various condition types from a unified model.

The main contributions of this paper are as follows:

- We present SCALAR, a controllable generation method based on VAR that introduces a Scale-wise Conditional Decoding mechanism, explicitly designed to align with the next-scale prediction nature of VAR models.
- A Unified Control Alignment process is introduced to SCALAR-Uni, enabling the various condition semantics guidance in controllable visual autoregressive learning.
- Extensive experiments on ImageNet demonstrate that **SCALAR** and SCALAR-Uni achieve exceptional generation quality and conditional consistency across all five conditional generation tasks (e.g., FID on Canny: **2.14** vs. 7.85; Depth: **3.09** vs. 4.19).

# Related Work

## Image Generation

Diffusion models (Rombach et al. 2022; Zhang et al. 2025; Peebles and Xie 2023) have achieved strong image and video generation performance but remain computationally expensive, motivating alternatives. In contrast, AR models treat image synthesis as sequence modeling, predicting tokens step by step. Early works (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016) focused on pixel-level generation. Building on the success of large language models (Touvron et al. 2023a; Achiam et al. 2023), emerging approaches adopt discrete quantizers such as VQ-VAE (Van Den Oord, Vinyals et al. 2017) and VQ-GAN (Esser, Rombach, and Ommer 2021) to convert image patches into discrete token indices, enabling next-token prediction over visual token sequences. Recent methods such as LlamaGen (Sun et al. 2024), Open-MAGVIT2 (Luo et al. 2024), and AiM (Li et al. 2024a) leverage LLaMA (Touvron et al. 2023b) or Mamba (Gu and Dao 2023) backbones. Among them, Visual Autoregressive Modeling (VAR) introduces a scalable next-scale prediction mechanism, which differs from raster-scan AR methods. VAR-based models for class-to-image and text-to-image tasks (Tian et al. 2024; Ma et al. 2024; Tang et al. 2024; Han et al. 2025; Zhuang et al. 2025) achieve image synthesis performance comparable to state-of-the-art diffusion models while offering significant computational efficiency.

## Controllable Image Generation

Following the success of diffusion models, diffusion-based controllable generation methods have been thoroughly studied (Mou et al. 2024a; Qin et al. 2023; Li et al. 2024b, 2025d; Wu et al. 2025; Luan et al. 2025; Li et al. 2025c,b,a). In contrast, conditional generation for AR models has been far less explored. Existing works on controllable autoregressive learning are typically classified according to the guidance injection phase in autoregressive generation: pre-filling phase or decoding phase. Pre-filling-phase methods (Li et al. 2024c; Qu et al. 2025) fill the control encodings into the initial sequence from the start. This approach increases the token sequence length, inflating the computational load, while the joint modeling scheme in ControlVAR (Li et al. 2024c) can disrupt the powerful capabilities of the pretrained backbone. Decoding-phase methods (Li et al. 2024d), which continuously inject control encodings throughout the step-by-step decoding, are simpler and more flexible. However, CAR (Yao et al. 2024) adopts parallel conditional decoding branches inspired by ControlNet, which introduces huge complexities when autoregressive decoding. Notably, although VAR models generally outperform traditional AR models in image synthesis quality, these VAR-based controllable methods (Li et al. 2024c; Yao et al. 2024) still underperform compared to raster-scan AR-based methods (Li et al. 2024d) in terms of control consistency and image quality. This gap indicates that the potential of VAR in controllable generation is yet to be fully explored. In this paper, we aim to develop a general and efficient method for controllable image generation based on VAR.

# SCALAR

## Preliminary

Building on the success of large language models, autoregressive models use discrete quantizers (e.g., VQ-VAE) to map image patches to tokens for next-step prediction, while VAR introduces a scalable next-scale prediction across resolutions. VAR-based class-to-image (c2i) and text-to-image (t2i) models achieve image synthesis performance comparable to state-of-the-art diffusion models, at a lower computational cost.

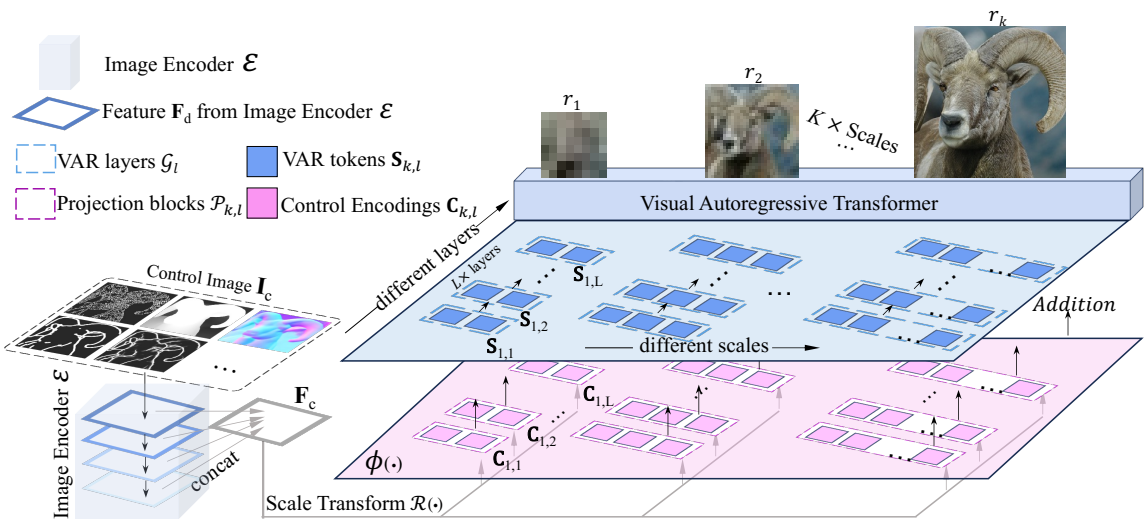


Figure 2: The framework of our **SCALAR** applies a next-scale paradigm adapted for VAR to design a Scale-wise Conditional Decoding mechanism. The feature $\mathbf{F}_c$ is obtained by concatenating four features $\mathbf{F}_d$ extracted by the Image Encoder $\mathcal{E}$.

The core idea of the VAR model lies in next-scale prediction, which contrasts with traditional raster-scan AR models based on next-token prediction. While conventional AR models generate images token by token along a flattened pixel sequence, this formulation suffers from mathematical inconsistencies and often leads to structural degradation, especially in highly structured images. In contrast, VAR avoids these issues by operating at the level of token maps, predicting the image progressively across multiple scales. Starting from a coarse $1 \times 1$ token map $r_1$, it autoregressively generates a sequence of higher-resolution token maps $(r_2, \ldots, r_K)$, each representing a finer level of detail. The overall generation process is formulated as:

$$p(r_1, r_2, \ldots, r_K) = \prod_{k=1}^{K} p(r_k \mid r_{<k}), \tag{1}$$

where $r_k \in [V]^{h_k \times w_k}$ represents the token map at scale $k$, with dimensions $h_k$ and $w_k$, conditioned on previous maps $r_{<k}$. Each token in $r_k$ is an index from the VQVAE codebook $V$, which is trained through multi-scale quantization and shared across scales. A standard cross-entropy loss is used to supervise VAR, defined as:

$$\mathcal{L}_{CE} = \mathbb{E}_{r_k \sim p(r_k)} \left[ -\log p_\theta(r_k \mid r_{<k}) \right]. \tag{2}$$

The loss is applied at each scale to train the model to predict finer token maps conditioned on coarser ones.

## Scale-wise Conditional Decoding

The generation process in autoregressive models is divided into two phases: pre-filling and decoding. In SCALAR, we diverge from the conditional pre-filling way (Li et al. 2024c; Qu et al. 2025), which relies on a distinct initial conditioning filling step. We adopt a simpler conditional decoding strategy (Li et al. 2024d; Yao et al. 2024). This design integrates the control signal directly and continuously throughout the decoding process. The scale-autoregressive nature of VARs poses distinct requirements for how control signals are applied across the generative hierarchy; guidance must be present and adapted at each scale. SCALAR achieves this by injecting the control signal into a predefined subset of layers, indexed by $l \in \mathcal{S}$. Formally, the operation at each layer is expressed as:

$$\mathbf{S}'_{k,l} = \mathcal{G}_l(\mathbf{S}_{k,l} + \mathbf{C}_{k,l}), \tag{3}$$

where $\mathcal{G}_l$ represents the $l$-th layer of the GPT-style decoder transformer, $\mathbf{S}_{k,l}$ and $\mathbf{S}'_{k,l}$ denote the input and output sequence of image tokens at layer $l$ for scale $k$, and $\mathbf{C}_{k,l}$ is the corresponding control signal encoding sequence injected at that same layer and scale. The set $\mathcal{S} \subseteq \{0, \ldots, L-1\}$ contains the indices of the layers targeted for injection, where $L$ is the total depth of the decoder $\mathcal{G}$.

**Control Signal Encoding $\mathbf{C}_{k,l}$.** We employ a pretrained vision foundation model (Oquab et al. 2023) as our universal control signal encoder. Its inherent understanding of rich visual semantics, learned via large-scale self-supervision, provides a powerful and robust feature extractor for diverse control conditions. To leverage the hierarchical features within the encoder (Bolya et al. 2025a), the control representation $\mathbf{F}_c$ is formed by concatenating features from the $i$-th layers of the encoder $\mathcal{E}$. It can be expressed as:

$$\mathbf{F}_c = \mathcal{C}\left( \underset{i \in \mathcal{I}}{\mathcal{E}_i}(\mathbf{I}_c) \right), \tag{4}$$

where $\mathcal{E}_i(\cdot)$ is the feature from the $i$-th layer of the encoder, $\mathcal{C}(\cdot)$ represents the concatenation operation and $\mathcal{I}$ is the set of indices for the selected layers. The resulting scale-agnostic feature $\mathbf{F}_c$ is then tailored for injection into scale $k$

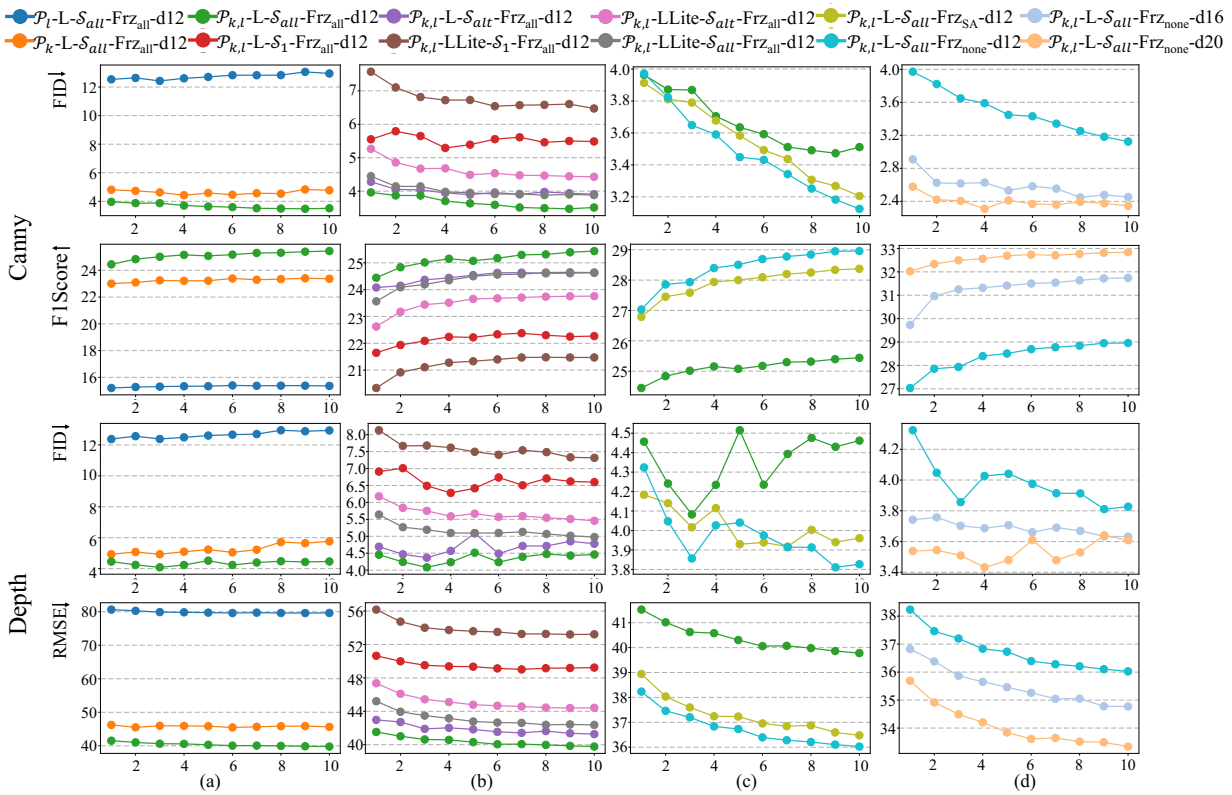


Figure 3: (a) Comparison of parameter sharing for projection blocks ($\mathcal{P}_{k,l}$, $\mathcal{P}_k$, and $\mathcal{P}_l$). (b) Comparison of various injection layers set ($\mathcal{S}_1$, $\mathcal{S}_{\text{alt}}$, and $\mathcal{S}_{\text{all}}$) with different structures of projection block (**Linear** and **LinearLite**). (c) Comparison of different parameter-efficient training strategies (Frz$_{\text{none}}$, Frz$_{\text{SA}}$, and Frz$_{\text{all}}$). (d) Impacts of scaling up the depth of VAR backbone (VAR-d12, d16, and d20).

and layer $l$ of our Scale-wise Conditional Decoding process:

$$\mathbf{C}_{k,l} = \phi_{k,l}(\mathbf{F}_c) = \mathcal{P}_{k,l}(\mathcal{R}_{h_k,w_k}(\mathbf{F}_c)), \tag{5}$$

where $\mathcal{R}_{h_k,w_k}$ denotes transforming the features to the target spatial size $(h_k, w_k)$ of scale $k$. Subsequently, a conditional projection block $\mathcal{P}_{k,l}$ maps to the final control signal encoding $\mathbf{C}_{k,l}$.

## Exploring Controlled Architectural Design

We explore several strategies to determine the optimal architecture for SCALAR. Our primary goal is to identify the most effective designs for encoding and injecting control signals into the VAR's generative process. We focus our experiments on the following key architectural observations:

- **Parameter Sharing of $\mathcal{P}_{k,l}$.** The control signal is injected via projection blocks. We investigate different parameter-sharing strategies for these layers. Specifically, we explore whether the projection weights should be: (i) unique for each transformer layer at each scale, marked as $\mathcal{P}_{k,l}$, (ii) shared across all layers, marked as $\mathcal{P}_k$, or (iii) shared across all scales, marked as $\mathcal{P}_l$.
- **Projection Blocks $\mathcal{P}_{k,l}$ and Injection Layers Set $\mathcal{S}$.** By default, $\mathcal{P}_{k,l}$ is a single linear layer, which can be parameter-intensive, especially when the injection set $\mathcal{S}$ is large. We explore two main aspects:
  - **Structure of $\mathcal{P}_{k,l}$**: We compare (i) **Linear**: a standard linear layer against (ii) **LinearLite**: a more parameter-efficient bottleneck structure, which consists of two linear layers that first squeeze the channel dimension and then expand it back.
  - **Injection Set $\mathcal{S}$**: We investigate how the density of control signal injection affects performance by testing three configurations for $\mathcal{S}$: (i) injecting only into the first layer $\mathcal{S}_1 = \{1\}$, (ii) injecting into alternating layers $\mathcal{S}_{\text{alt}} = \{1, 3, 5, \ldots, L-1\}$, and (iii) injecting into all layers $\mathcal{S}_{\text{all}} = \{1, 2, 3, \ldots, L\}$.
- **Parameter-Efficient Training**. We analyze various training strategies to evaluate the trade-off between performance and the number of trainable parameters. Specifically, we compare: (i) Frz$_{\text{none}}$: fine-tuning all VAR parameters, (ii) Frz$_{\text{SA}}$: freeze all self-attention layers of VAR, and (iii) Frz$_{\text{all}}$: freeze all VAR parameters.

For all comparisons above, we train VAR-d12 models on ImageNet $256 \times 256$ with batch size 512. We train models with the above changes and compare them on Fréchet Inception Distance (FID), Inception Score (IS), Root Mean Square Error (RMSE, conditioned on Depth), and F1-Score (conditioned on Canny). As shown in Fig. 3(a) and (b), the optimal configuration employs projection blocks with scale-

| Type | Method | Model | Canny | | | Depth | | | Normal | | | HED | | | Sketch | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FID | IS | F1 | FID | IS | RMSE | FID | IS | RMSE | FID | IS | SSIM | FID | IS | F1 |
| Diff. | T2IAdapter | - | ~10.2 | ~157 | - | ~9.9 | ~134 | - | ~9.5 | ~143 | - | ~9.3 | ~142 | - | ~16.2 | ~156 | - |
| | ControlNet | - | ~11.6 | ~173 | - | ~9.2 | ~150 | - | ~8.9 | ~155 | - | ~8.6 | ~150 | - | ~15.3 | ~163 | - |
| AR | ControlAR | AiM-L | 9.66 | - | 30.4 | 7.39 | - | 35.0 | | - | | | - | | | - | |
| | | LG-B | 10.64 | - | 34.2 | 6.67 | - | 32.4 | | - | | | - | | | - | |
| | | LG-L | 7.69 | - | 34.9 | 4.19 | - | 31.1 | | - | | | - | | | - | |
| VAR | ControlVAR | d12 | ~35.2 | ~44 | - | ~26.7 | 52 | - | ~25.3 | ~55 | - | | - | | | - | |
| | | d16 | ~16.2 | ~80 | - | ~13.8 | 92 | - | ~14.2 | ~89 | - | | - | | | - | |
| | | d20 | ~13.0 | ~94 | - | ~13.4 | 98 | - | ~12.8 | ~100 | - | | - | | | - | |
| | | d24 | ~15.7 | ~100 | - | ~12.5 | 125 | - | ~11.8 | ~123 | - | | - | | | - | |
| | | d30 | 7.85 | 160.0 | - | 6.50 | 180.5 | - | 6.20 | 172.0 | - | | - | | | - | |
| | CAR | d16 | ~12.8 | ~85 | - | ~10.8 | ~95 | - | ~11.0 | ~98 | - | ~9.8 | ~102 | - | ~13.2 | ~83 | - |
| | | d20 | ~10.2 | ~125 | - | ~8.0 | ~135 | - | ~8.8 | ~138 | - | ~7.2 | ~144 | - | ~11.2 | ~118 | - |
| | | d24 | ~9.0 | ~155 | - | ~7.0 | ~165 | - | ~7.5 | ~168 | - | ~6.5 | ~176 | - | ~10.0 | ~143 | - |
| | | d30 | 8.30 | 167.3 | - | 6.90 | 178.6 | - | 6.60 | 175.9 | - | 5.60 | 182.2 | - | 10.20 | 161.6 | - |
| | SCALAR (Ours) | d12 | 3.12 | 191.3 | 29.0 | 3.83 | 233.3 | 36.03 | 3.76 | 225.7 | 28.14 | 2.62 | 189.1 | 74.93 | 4.55 | 229.6 | 76.6 |
| | | d16 | 2.45 | 237.7 | 31.7 | 3.63 | 285.8 | 34.77 | 3.70 | 279.5 | 27.64 | 1.97 | 222.5 | 76.37 | 4.51 | 292.4 | 77.0 |
| | | d20 | 2.34 | 254.3 | 32.8 | 3.61 | 301.3 | 33.34 | 3.51 | 300.9 | 27.57 | 1.81 | 240.4 | 76.74 | 4.22 | 312.7 | 77.4 |
| | | d24 | 2.14 | 261.8 | 33.1 | 3.09 | 306.9 | 33.13 | 3.09 | 307.2 | 27.45 | 1.72 | 244.2 | 76.98 | 3.57 | 315.5 | 77.6 |
| | SCALAR-Uni (Ours) | d12 | 3.31 | 207.4 | 27.5 | 4.21 | 231.5 | 38.57 | 4.05 | 222.5 | 29.74 | 2.82 | 218.4 | 71.10 | 4.75 | 222.9 | 75.4 |
| | | d16 | 2.78 | 262.4 | 30.8 | 3.73 | 294.1 | 37.24 | 4.03 | 288.6 | 29.17 | 2.53 | 263.9 | 72.73 | 4.58 | 297.8 | 75.7 |
| | | d20 | 2.64 | 276.9 | 31.7 | 3.69 | 310.1 | 37.12 | 3.56 | 308.5 | 29.48 | 2.37 | 272.6 | 72.82 | 4.05 | 311.8 | 75.8 |

Table 1: Quantitative results for conditional image generation on ImageNet. LG in "model" denotes LlamaGen, and F1 denotes F1-Score. Values marked with ~ are estimated from reported histograms. Metrics: FID↓, IS↑, F1-Score↑, RMSE↓, SSIM↑.

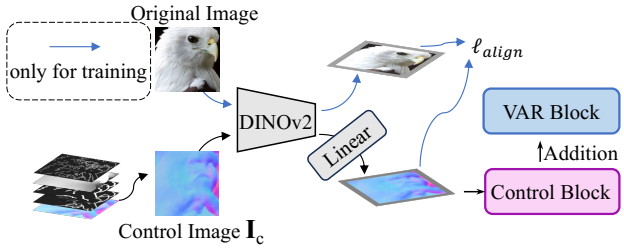


Figure 4: Our SCALAR-Uni, a unified multi-condition control method. During training, control images are randomly sampled with equal probability.

wise parameters for each block ($\mathcal{P}_{k,l}$), uses a standard linear layer (**Linear**), and injects the control signal into all layers ($\mathcal{S}_{\text{all}}$). Fig. 3(c) shows that freezing backbone parameters causes a significant reduction of control performance. In Fig. 3(d), scaling up depth of VAR helps performance. Accordingly, we adopt the best combination of settings ($\mathcal{P}_{k,l}$-**Linear**-$\mathcal{S}_{\text{all}}$-Frz$_{\text{none}}$) for our final SCALAR model to achieve the best performance. Further analysis details about Fig. 3 are provided in the **Supplementary Material**.

## Unified Control Alignment

Thanks to the architectural simplicity of SCALAR, extending it to manage multiple conditions concurrently only requires addressing a key challenge: the control features extracted for different modalities (e.g., Canny, Depth) reside in distinct and potentially incompatible feature spaces. As shown in Fig. 4, we propose a unified version, **SCALAR-Uni**. The core idea is to map disparate control features into a common, modality-agnostic latent space. We utilize the image feature space itself as the target for alignment, as it offers a rich and universal representation of visual concepts. We enforce alignment by introducing an auxiliary loss during training, minimizing the L2 distance between the projected controls and the corresponding image features:

$$\begin{aligned}\mathcal{L}_{\text{align}} &= \|\mathcal{F}_{\text{align}}(\mathbf{F}_c) - \mathbf{F}_{\text{img}}\|_2^2 \\ &= \left\|\mathcal{F}_{\text{align}}(\mathbf{F}_c) - \mathcal{C}\left(\underset{i\in\mathcal{I}}{\mathcal{E}_i}(\mathbf{I}_{\text{img}})\right)\right\|_2^2 \end{aligned} \tag{6}$$

where $\mathbf{I}_{\text{img}}$ represents the corresponding image for control signal, $\mathcal{F}_{\text{align}}$ represents a lightweight alignment module, implemented as a linear layer, learning to project the initial control representation $\mathbf{F}_c$ into this shared image feature space. The total training loss of SCALAR-Uni can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{align}}, \tag{7}$$

where $\lambda$ is a constant that regulates the alignment loss.

# Results

## Experimental Setup

**Dataset.** We conduct experiments on the ImageNet-256 (Deng et al. 2009), using 50K images from the validation set for evaluation. To assess the controllable generation capability of our model, we consider five types of conditions: Canny (Canny 1986), Depth (Ranftl et al. 2020), Normal (Vasiljevic et al. 2019), HED (Xie and Tu 2015), and Sketch (Su et al. 2021). **Training Details.** We extract four features from Image Encoder $\mathcal{E}$ at the layers specified by the index set $\mathcal{I} = \{d - 1 - k \cdot \lfloor d/4 \rfloor \mid k \in \{0, 1, 2, 3\}\}$. We use DINOv2 (Oquab et al. 2023) as the Image Encoder for its strong representations and scalability. We use pretrained

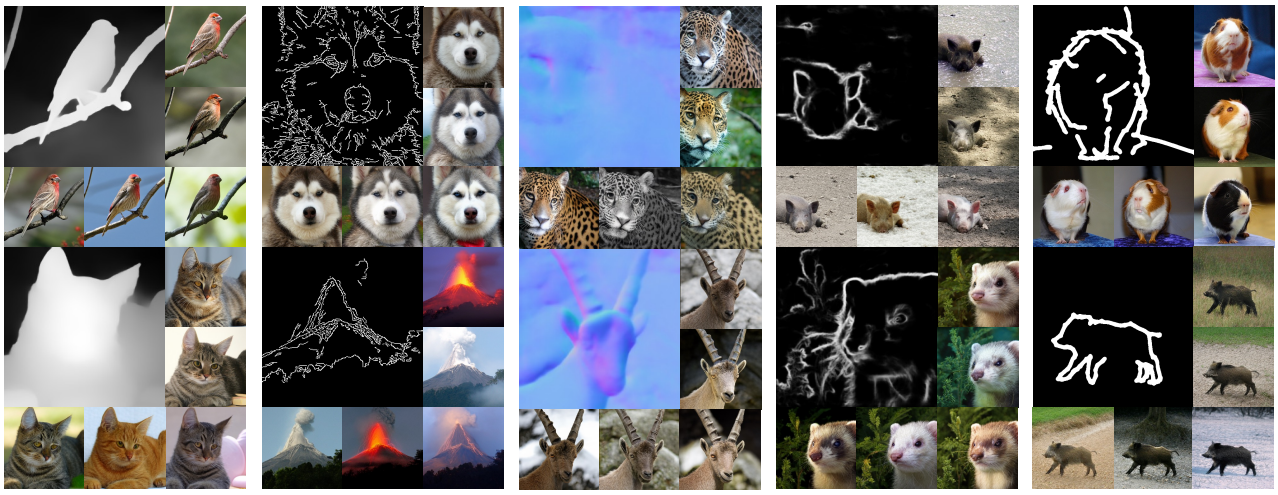

Figure 5: Visual results generated by SCALAR.

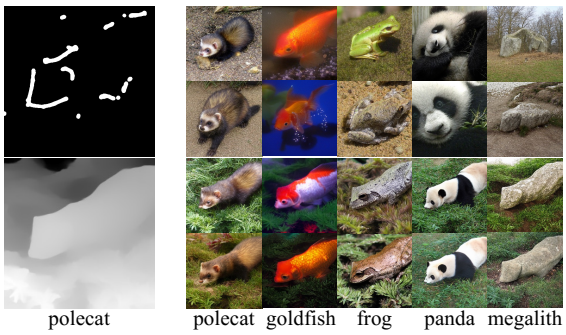


Figure 6: Results of controllable generation with SCALAR by changing class labels (e.g., "polecat" → "goldfish").

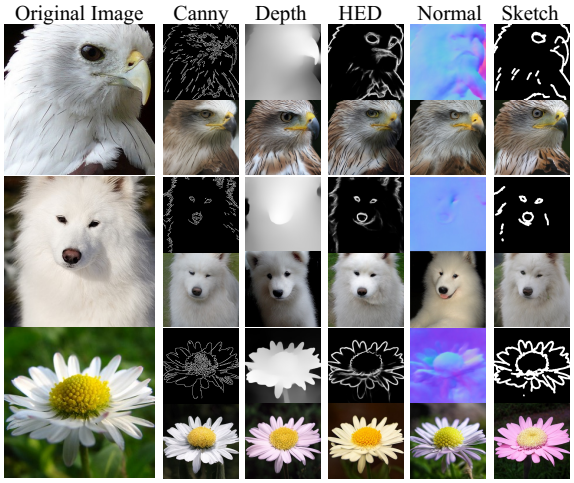


Figure 7: Visual results generated by unified multi-condition method SCALAR-Uni under varying control types.

VARs with depths of 12, 16, 20, and 24, adopting DINOv2-S for 12 and DINOv2-B otherwise. Following (Zhang, Rao, and Agrawala 2023b), the control block is zero-initialized. Our SCALAR and SCALAR-Uni are trained for 10 epochs with the AdamW optimizer on 8 H20 GPUs. For SCALAR-Uni, we sample five control types with equal probability.

## Experimental Results

We evaluate both conditional consistency and image generation quality of SCALAR and SCALAR-Uni, and we compare them with existing controllable generation methods, based on diffusion models (Zhang, Rao, and Agrawala 2023b; Mou et al. 2024b), raster-scan AR models (Li et al. 2024d), and VAR methods (Li et al. 2024c; Yao et al. 2024).

**Results for SCALAR.** Results show that SCALAR outperforms existing methods in generation quality, achieving superior FID and IS scores. Notably, even SCALAR-d12 achieves better FID compared to ControlAR with LlamaGen-L, while VAR-d12 uses only **50.1%** of the parameters of LlamaGen-L (e.g., FID on Canny: **3.12** vs. 7.85; Depth: **3.83** vs. 4.19). In terms of conditional consistency, SCALAR also demonstrates competitive performance against other State-of-the-Art (SoTA) methods. Moreover, as the depth is scaled up, both the generation quality and the conditional consistency of SCALAR steadily improve. Qualitative results are shown in Fig. 5. Besides, SCALAR also demonstrates strong generalization by synthesizing an object of the target class while adhering to the spatial structure of the source control image, even when the label and control structure are inconsistent, as illustrated in Fig. 6.

**Results for SCALAR-Uni.** Compared to other SoTA methods, SCALAR-Uni demonstrates clear advantages in image generation quality while maintaining high conditional consistency. When compared to SCALAR, SCALAR-Uni shows a slight drop in both generation quality and consistency, which may be attributed to the need to accommodate multiple control conditions, thereby reducing the effective training data per condition. Overall, SCALAR-Uni exhibits strong performance and generalization as a unified frame-

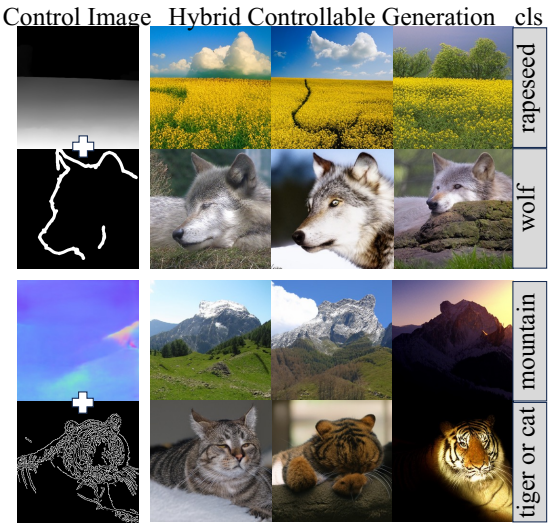


Figure 8: Results of zero-shot hybrid controllable generation with **SCALAR-Uni**, combining two control types (e.g., Depth+Sketch) to generate images with both characteristics.

work for multi-condition controllable generation. More visualization details can be found in Fig. 7.

**Zero-shot Hybrid Controllable Synthesis.** SCALAR-Uni is tested. We consider two different categories of control images as inputs (e.g., Depth+Sketch). In addition to the corresponding control images, the class conditions of both inputs are also fed into SCALAR-Uni. No retraining is conducted. As shown in Fig. 8, SCALAR-Uni successfully combines different control types and generates high-quality images with the characteristics of both controls. It demonstrates the strong generalization ability of SCALAR-Uni, where the Unified Control Alignment maps diverse control features into a common, modality-agnostic latent space.

## Ablation Study

**Ablations on Image Encoder $\mathcal{E}$.** In Tab. 2, we conduct experiments using different image encoders (or pretraining schemes). Unlike previous approaches (Li et al. 2024d; Yao et al. 2024), SCALAR and SCALAR-Uni adopt frozen image encoders to preserve the robust features obtained from large-scale self-supervised pretraining. Inspired by (Bolya et al. 2025b; Ye et al. 2025), we extract multi-layer features including intermediate layers, instead of relying solely on the final output, which improves image generation quality while maintaining control consistency. As shown in Tab. 2 (a) to (e), DINOv2 outperforms other pretrained vision models such as ViT (Dosovitskiy 2020) and SAM (Kirillov et al. 2023). Furthermore, comparing (b), (c), (a), and (e) indicates that increasing the scale of the image encoder significantly boosts performance under the same backbone. Considering the trade-off between parameter size and effectiveness, we use DINOv2-S for the d12 architecture and DINOv2-B for all other settings. The comparison from (e) to (h) demonstrates that deeper VAR leads to further improvements in both generation quality and control accuracy, suggesting that deeper models can better exploit visual representations.

| Idx | Image Encoder $\mathcal{E}$ | Para. of $\mathcal{E}$ | Model | Canny FID↓ | Canny F1↑ | Depth FID↓ | Depth RMSE↓ |
|---|---|---|---|---|---|---|---|
| (a) | DINOv2-S | 22.1M | d12 | 3.12 | 29.0 | 3.83 | 36.03 |
| (b) | ViT-S | 21.8M | d12 | 4.34 | 24.2 | 5.01 | 42.00 |
| (c) | ViT-B | 86.4M | d12 | 4.23 | 26.0 | 4.83 | 40.77 |
| (d) | SAM-B | 89.6M | d12 | 6.13 | 28.4 | 5.72 | 40.41 |
| (e) | DINOv2-B | 86.6M | d12 | 2.95 | 29.4 | 3.68 | 35.50 |
| (f) | DINOv2-B | 86.6M | d16 | 2.45 | 31.7 | 3.63 | 34.77 |
| (g) | DINOv2-B | 86.6M | d20 | 2.34 | 32.8 | 3.61 | 33.34 |
| (h) | DINOv2-B | 86.6M | d24 | 2.14 | 33.1 | 3.09 | 33.13 |

Table 2: Ablation of Image Encoder $\mathcal{E}$ and the depth of our SCALAR.

| Model | Canny FID↓ | Canny F1↑ | Depth FID↓ | Depth RMSE↓ |
|---|---|---|---|---|
| **SCALAR-Uni-d12** | **3.31** | **27.5** | **4.21** | **38.57** |
| ↳ *w/o* $\mathcal{L}_{align}$ *in eq. 6* | 3.73 | 26.4 | 4.40 | 40.79 |
| **SCALAR-Uni-d16** | **2.78** | **30.8** | **3.73** | **37.24** |
| ↳ *w/o* $\mathcal{L}_{align}$ *in eq. 6* | 2.88 | 30.2 | 4.09 | 37.85 |
| **SCALAR-Uni-d20** | **2.64** | **31.7** | **3.69** | **37.12** |
| ↳ *w/o* $\mathcal{L}_{align}$ *in eq. 6* | 2.84 | 31.5 | 3.74 | 36.88 |

Table 3: Ablations on SCALAR-Uni regarding the unified control alignment $\mathcal{L}_{align}$.

**Ablations on Unified Control Alignment.** We conduct ablations on SCALAR-Uni to validate the effectiveness of Unified Control Alignment on Canny and Depth. As shown in Tab. 3, thanks to the strong general visual representation capability of DINOv2, *SCALAR* can effectively capture the differences across modalities and achieve preliminary alignment of multi-modal control signals. After incorporating Unified Control Alignment, **SCALAR-Uni** consistently outperforms *SCALAR* across different network depths in both image generation quality and control consistency. Results in Tab. 3 demonstrate that **SCALAR-Uni**, constructed by integrating *SCALAR* with Unified Control Alignment, offers a simple and effective solution for controllable image generation under diverse conditions.

## Conclusion and Future Work

**Conclusion.** In this work, we propose SCALAR, a controllable generation method based on VAR that introduces a Scale-wise Conditional Decoding mechanism adapted for the hierarchical nature of VAR models. By leveraging a pretrained image encoder to extract semantically rich control features and injecting them into scale-specific layers, SCALAR enables persistent and structured guidance throughout the generation process. Our extension, SCALAR-Uni, further supports unified multi-conditional control. We hope our findings will inspire further research on controllable generation within VAR models.

**Future Work.** (1) The spatial transform to align with each scale is worth further study. (2) SCALAR is currently evaluated only on ImageNet. We plan to extend to t2i generation.

## References


Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bolya, D.; Huang, P.-Y.; Sun, P.; Cho, J. H.; Madotto, A.; Wei, C.; Ma, T.; Zhi, J.; Rajasegaran, J.; Rasheed, H.; et al. 2025a. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*.

Bolya, D.; Huang, P.-Y.; Sun, P.; Cho, J. H.; Madotto, A.; Wei, C.; Ma, T.; Zhi, J.; Rajasegaran, J.; Rasheed, H.; et al. 2025b. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2025. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15733–15744.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, H.; Jiang, L.; Xiao, X.; Wang, T.; Yi, H.; Wu, B.; and Cai, D. 2025a. MagicID: Hybrid Preference Optimization for ID-Consistent and Dynamic-Preserved Video Customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12737–12746.

Li, H.; Qiu, H.; Zhang, S.; Wang, X.; Wei, Y.; Li, Z.; Zhang, Y.; Wu, B.; and Cai, D. 2025b. PersonalVideo: High ID-Fidelity Video Customization without Dynamic and Semantic Degradation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 19406–19416.

Li, H.; Xu, J.; Cheng, K.; Wang, L.; Bi, N.; Wu, B.; la Torre, F. D.; and Cai, D. 2025c. PersonalView: Multi-View Consistent Human Image Customization via In-Context Learning.

Li, H.; Yang, J.; Wang, K.; Qiu, X.; Chou, Y.; Li, X.; and Li, G. 2024a. Scalable Autoregressive Image Generation with Mamba. *arXiv preprint arXiv:2408.12245*.

Li, M.; Yang, T.; Kuang, H.; Wu, J.; Wang, Z.; Xiao, X.; and Chen, C. 2024b. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. *arXiv preprint arXiv:2404.07987*.

Li, X.; Qiu, K.; Chen, H.; Kuen, J.; Lin, Z.; Singh, R.; and Raj, B. 2024c. ControlVAR: Exploring Controllable Visual Autoregressive Modeling. *arXiv preprint arXiv:2406.09750*.

Li, Y.; Zhang, S.; Chen, Y.; Li, B.; Zhang, Y.; and Du, X. 2025d. SpotDiff: Spotting and Disentangling Interference in Feature Space for Subject-Preserving Image Generation. *arXiv preprint arXiv:2510.07340*.

Li, Z.; Cheng, T.; Chen, S.; Sun, P.; Shen, H.; Ran, L.; Chen, X.; Liu, W.; and Wang, X. 2024d. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*.

Luan, S.; Wang, Z.; Shen, L.; Gu, Z.; Wu, C.; and Tao, D. 2025. Dynamic neural fortresses: An adaptive shield for model extraction defense. In *The Thirteenth International Conference on Learning Representations*.

Luo, Z.; Shi, F.; Ge, Y.; Yang, Y.; Wang, L.; and Shan, Y. 2024. Open-MAGVIT2: An Open-Source Project Toward Democratizing Auto-regressive Visual Generation. *arXiv preprint arXiv:2409.04410*.

Ma, X.; Zhou, M.; Liang, T.; Bai, Y.; Zhao, T.; Chen, H.; and Jin, Y. 2024. Star: Scale-wise text-to-image generation via auto-regressive representations. *arXiv preprint arXiv:2406.10797*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024a. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024b. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J. C.; Xiong, C.; Savarese, S.; et al. 2023. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*.

Qu, Y.; Yuan, K.; Hao, J.; Zhao, K.; Xie, Q.; Sun, M.; and Zhou, C. 2025. Visual Autoregressive Modeling for Image Super-Resolution. *arXiv preprint arXiv:2501.18993*.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Su, Z.; Liu, W.; Yu, Z.; Hu, D.; Liao, Q.; Tian, Q.; Pietikäinen, M.; and Liu, L. 2021. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5117–5127.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv preprint arXiv:2406.06525*.

Tang, H.; Wu, Y.; Yang, S.; Xie, E.; Chen, J.; Chen, J.; Zhang, Z.; Cai, H.; Lu, Y.; and Han, S. 2024. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Van Den Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *International conference on machine learning*, 1747–1756. PMLR.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Vasiljevic, I.; Kolkin, N.; Zhang, S.; Luo, R.; Wang, H.; Dai, F. Z.; Daniele, A. F.; Mostajabi, M.; Basart, S.; Walter, M. R.; et al. 2019. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*.

Wu, C.; Wang, Z.; Xie, K.; Devulapally, N. K.; Lokhande, V. S.; and Gao, M. 2025. Model-Agnostic Gender Bias Control for Text-to-Image Generation via Sparse Autoencoder. arXiv:2507.20973.

Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, 1395–1403.

Yao, Z.; Li, J.; Zhou, Y.; Liu, Y.; Jiang, X.; Wang, C.; Zheng, F.; Zou, Y.; and Li, L. 2024. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*.

Ye, D.; Fan, C.; Huang, Z.; Luo, C.; Li, J.; Yu, S.; and Liu, X. 2025. BiggerGait: Unlocking Gait Recognition with Layer-wise Representations from Large Vision Models. *arXiv preprint arXiv:2505.18132*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, L.; Rao, A.; and Agrawala, M. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, N.; Li, Y.; Du, D.; Chong, Z.; Sun, Z.; Zeng, J.; Dai, Y.; Xie, Z.; Zhu, H.; and Han, X. 2025. Robust-MVTON: Learning Cross-Pose Feature Alignment and Fusion for Robust Multi-View Virtual Try-On. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16029–16039.

Zhao, X.; Liu, B.; Liu, Q.; Shi, G.; and Wu, X.-M. 2023. Easygen: Easing multimodal generation with bidiffuser and llms. *arXiv preprint arXiv:2310.08949*.

Zhou, G.; Pan, H.; LeCun, Y.; and Pinto, L. 2024. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*.

Zhuang, X.; Xie, Y.; Deng, Y.; Liang, L.; Ru, J.; Yin, Y.; and Zou, Y. 2025. VARGPT: Unified Understanding and Generation in a Visual Autoregressive Multimodal Large Language Model. arXiv:2501.12327.