S2AR: DUAL-AXIS ROUTING FOR CONTINUOUS ATTRIBUTE CONTROL IN VISUAL AUTOREGRESSIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual autoregressive models have emerged as a powerful and scalable paradigm for image synthesis. However, continuous attribute control within this framework remains underexplored. To address this gap, we present S2AR, a training-free framework for continuous attribute control. Specifically, we formulate attribute control as a dual-axis routing problem across spatial and scale dimensions. S2AR first identifies target visual tokens to construct a spatial localization map, and subsequently determines scale-wise routing weights based on structural sensitivity. Guided by these routing weights, a single control scalar jointly modulates cross-attention and global conditioning, thereby continuously steering attribute intensity. On our constructed Continuous Attribute Control Benchmark, S2AR consistently outperforms existing training-free and LoRA-based methods while faithfully preserving identity and background details.

1 INTRODUCTION

Recent text-to-image (T2I) models have achieved remarkable visual quality (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022; Yu et al., 2022). At the same time, practical applications increasingly demand finer control: users may wish not only to specify image content, but also to continuously adjust the strength of a particular attribute (Gandikota et al., 2024; Baumann et al., 2025; Chen et al., 2026). For example, a user may want to turn a subtle smile into a broad one while preserving the person's identity, pose, and background. Meeting this need requires continuous attribute control that smoothly changes the target attribute while retaining content unrelated to the edit (Parihar et al., 2024; Chiu et al., 2026).

Existing work has explored several approaches to attribute control. Early generative adversarial network (GAN) methods edited attributes such as age, expression, and lighting by discovering interpretable directions in latent space (Shen et al., 2020; Härkönen et al., 2020; Abdal et al., 2021). Diffusion models subsequently enabled more flexible slider-based control: Concept Sliders (Gandikota et al., 2024) train Low-Rank Adaptation (LoRA) modules, Text Sliders (Chiu et al., 2026) identify text-encoder directions that transfer across backbones, and Prompt Sliders (Sridhar & Vasconcelos, 2024) optimize concept embeddings. While these methods demonstrate the feasibility of continuously varying semantic attributes, they still require multi-step denoising, making repeated exploration of attribute strengths computationally expensive (Song et al., 2020; Song & Dhariwal, 2024; Lu & Song, 2025). More recently, visual autoregressive (VAR) models have combined competitive generation quality with efficient sampling (Tian et al., 2024; Tang et al.; Han et al., 2025). They provide a promising foundation for attribute control, but achieving precise, continuous edits while preserving image content in this framework remains an open challenge.

The underlying difficulty is that different scales in visual autoregressive generation serve different purposes. Coarse scales establish global composition and high-level semantics; subsequent scales, conditioned on earlier outputs, progressively add local structures and textures. Directly changing the prompt to edit an attribute implicitly assumes that the desired semantic change will leave the original generation trajectory otherwise intact. This assumption can fail: an early deviation may be inherited and amplified by later scales, changing untargeted content such as identity, pose, and background (Wang et al., 2025; Fang et al., 2026; Xia et al., 2026). Moreover, attributes have different

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

spatial requirements. We find that changing a facial expression depends on local details at intermediate or high resolutions, whereas adding fog or snow must affect the scene from earlier, coarser scales (Mao et al., 2026). A predetermined editing region therefore cannot reliably accommodate both local and global attributes. The central question is how to identify where and at which scales an attribute should be changed, relative to the original generation.

To address this challenge, we propose Spatial-Scale Attribute Routing (S2AR), a training-free framework for continuous attribute control. S2AR first compares model predictions under the original and attribute-modified prompts while conditioning on the same source image. It extracts directional prediction differences relative to the original generation and aggregates them across channels and scales to form a source-conditioned spatial localization map that localizes the target attribute. We then downsample and reconstruct this map on each scale's grid to assess whether that scale can represent the required spatial detail. This suppresses attribute injection at coarse scales that cannot resolve a local edit, while allowing global changes with broad spatial support to act earlier in generation. Finally, we scale the filtered prediction difference by a continuous factor α and inject it during generation. Attribute strength can thus vary continuously, while the locations and scales of the edit are selected according to the attribute's response.

We evaluate S2AR on our Continuous Attribute Control Benchmark (CACBench). The results show that S2AR produces smooth attribute transitions and better preserves subject identity, composition, and background details than direct prompt editing and existing training-free methods. Our contributions are threefold:

- We analyze cross-scale trajectory divergence in visual autoregressive models and identify the propagation of early semantic perturbations as an important cause of unintended content changes.
- We propose S2AR, a training-free framework that uses source-conditioned spatial localization maps to localize attribute changes and spatial-scale routing to inject continuously adjustable prediction differences at the appropriate locations and scales.
- We introduce CACBench and experimentally validate the effectiveness of S2AR in continuous attribute control and in preserving subject identity, composition, and background details.

2 Related Work

Continuous Attribute Control. Continuous attribute control aims to vary a target semantic property smoothly while preserving unrelated image content. Most existing work has focused on diffusion models (Gandikota et al., 2024; Sridhar & Vasconcelos, 2024; Baumann et al., 2025; Brack et al., 2023; Liu et al., 2023; Haas et al., 2024). Concept Sliders (Gandikota et al., 2024) train attribute-specific low-rank adapters using contrasting text descriptions and adjust the LoRA scaling factor at inference time to control attribute strength. Prompt Sliders (Sridhar & Vasconcelos, 2024) instead learn adjustable concept embeddings through textual inversion, while Baumann et al. (2025) identify token-level semantic directions in pretrained CLIP text embeddings to modulate high-level, subject-specific attributes. These studies establish the feasibility of continuous control through semantic directions, although constructing and applying those directions may depend on a particular model architecture or an additional learning process.

Subsequent work has placed greater emphasis on compositional control, training cost, and scalability (Gandikota et al., 2025; Zarei et al., 2026; Ye et al., 2026). CompSlider (Zhu et al., 2025) maps text prompts and independent slider values to conditional priors, enabling multiple attributes to be combined at different strengths. Text Slider (Chiu et al., 2026) identifies low-rank directions in a pretrained text encoder that can be reused across diffusion models sharing that encoder. FreeSliders (Ezra et al., 2025) estimates concept-specific semantic modifications at inference time, eliminating the need to train an adapter for each concept. Despite their increased flexibility, these methods primarily operate within the iterative generation process of diffusion models.

More recently, ATA (Yang et al., 2026) has extended continuous attribute control to visual autoregressive models. Given a reference image and paired text conditions, it learns attribute-specific semantic directions and uses token arithmetic to adjust and combine attributes. ATA demonstrates the potential of continuous control in autoregressive generation, but obtaining its semantic directions

requires a reference image and an additional direction-learning procedure, while control is primarily applied through text-token representations. In contrast, S2AR derives attribute changes directly from the prediction responses of a frozen model under the original and target prompts. It then routes those changes according to their spatial support and generation scale, without reference images or additional optimization.

Controllable Generation in Visual Autoregressive Models. As visual autoregressive models have advanced, recent work has explored how to introduce more precise control into token-wise or scale-wise generation (Shin et al., 2026; Xu et al., 2026a; Jin et al., 2026; Zhang et al., 2026). For spatially conditioned generation, ControlAR (Li et al., 2025) encodes conditions such as edges, depth, and segmentation into control tokens that guide subsequent image-token prediction. ControlVAR (Li et al., 2024) jointly models images and pixel-level conditions within a next-scale prediction framework. CAR (Yao et al., 2024) progressively refines conditional features and injects them into the corresponding generation stages. SCALAR (Xu et al., 2026b) further transforms control features extracted by a pretrained image encoder into scale-specific representations, aligning structural guidance with coarse-to-fine generation. These approaches improve how visual autoregressive models respond to external visual conditions, but primarily address generation from specified edges, depth maps, or segmentation maps rather than continuous adjustment of a textual attribute.

A more closely related direction is training-free, text-guided editing with visual autoregressive models. AREdit (Wang et al., 2025) uses source–target prediction differences to identify edit-relevant regions and selectively reassembles tokens across scales to preserve unmodified content. BitResEdit (Zhang et al., 2026), designed for bitwise-residual autoregressive models, combines bit-level semantic guidance with scale-wise residual-code composition for localized editing. These methods show that internal prediction differences and multiscale representations can help constrain the extent of an edit. They primarily study text-guided editing of a given image, however. How to inject continuously adjustable attribute changes during text-to-image generation while adapting to an attribute's spatial extent and scale requirements remains insufficiently explored. S2AR addresses this problem by jointly determining where, at which scales, and with what strength an attribute change is applied.

3 Problem Formulation and Preliminaries

Problem Formulation. We study training-free continuous attribute control with a frozen multi-scale autoregressive text-to-image model G_θ. Given a base prompt P, a target attribute description a, a control strength $\alpha \in [0, 1]$, and a random state ϵ, an inference-time controller $\mathcal{E}_\theta$ produces

$$X_\alpha = \mathcal{E}_\theta(P, a, \alpha; \epsilon). \tag{1}$$

We denote the image generated from the base prompt without attribute control by $X_0 = G_\theta(P; \epsilon)$. At zero control strength, the controller is required to reproduce this image: $\mathcal{E}_\theta(P, a, 0; \epsilon) = X_0$. We use the same random state ϵ across control strengths. As α increases, the target attribute should become progressively stronger, while non-target content, such as subject identity and scene layout, should remain consistent with X_0. The controller operates only at inference time, without updating model parameters or performing test-time optimization.

Multi-Scale Bitwise Autoregressive Generation. The backbone generates an image over K coarse-to-fine scales. At scale k, its visual representation is a binary code map $\mathbf{z}_k \in \{0, 1\}^{h_k \times w_k \times D}$, where h_k and w_k are the spatial dimensions and D is the number of bit channels at each position. Given a text condition $\mathbf{c}$ encoded from the prompt and the code maps generated at preceding scales, $\mathbf{z}_{<k} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{k-1}\}$, the frozen model predicts binary logits at the current scale:

$$\mathbf{L}_k = \mathcal{G}_{\theta,k}(\mathbf{c}, \mathbf{z}_{<k}). \tag{2}$$

For spatial position u and bit channel d, the difference between the logits for values 1 and 0 defines the log-odds

$$r_{k,u,d} = L_{k,u,d}(1) - L_{k,u,d}(0). \tag{3}$$

The corresponding probability of sampling a value of 1 is $\Pr(z_{k,u,d} = 1) = \sigma(r_{k,u,d})$, where σ is the sigmoid function. Once sampled, $\mathbf{z}_k$ becomes part of the conditioning history for subsequent scales. The code maps specify multi-scale latent residuals, which are accumulated and decoded into the final image. Thus, predictions at later scales depend on the codes sampled at earlier scales.

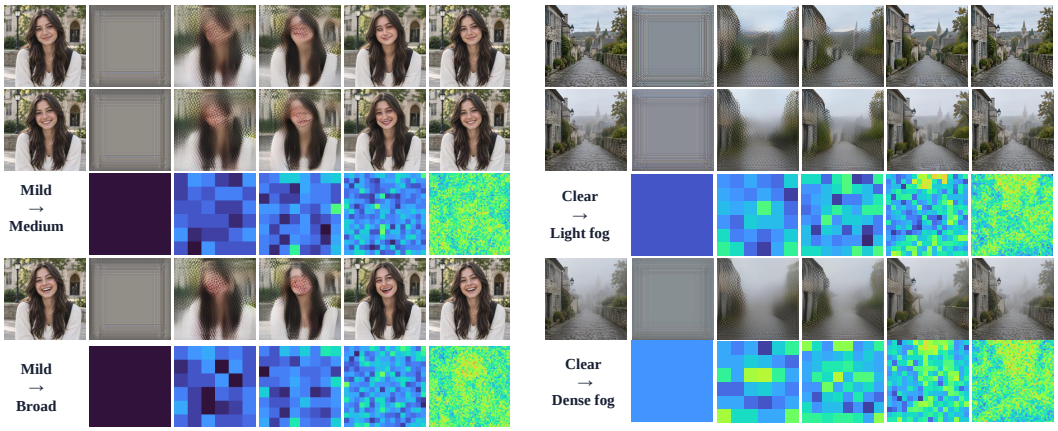


Figure 1: Exploratory analysis of two aligned image triplets varying smile intensity and fog density. Cumulative VAE decodes at selected scales and spatial bit-disagreement maps illustrate differences in the scale of onset and spatial extent of the two attributes.

4 Attribute Control as Dual-Axis Routing

VAR models generate images through a coarse-to-fine sequence of scales. Controlling an attribute therefore requires deciding both where its effect should appear in the image and at which generation scales it should be introduced. We formulate continuous attribute control as routing along these two axes: the spatial axis determines *where* to apply an editing signal, and the scale axis determines *at which scales* to introduce it.

Given a base prompt P and an edited prompt P^a describing attribute a, a direct approach is to use the difference between their conditions to modify model representations. Without spatial or scale selection, however, this intervention may also affect regions unrelated to the target attribute or alter the image composition. S2AR therefore comprises two complementary routing components. *Spatial attribute routing* uses model responses to construct a soft mask M that identifies where to edit. *Scale attribute routing* assigns a weight β_k to each generation scale and estimates how broadly the response is distributed, distinguishing local changes from scene-wide ones. During generation, these quantities jointly regulate the extent and strength of attribute injection.

Exploratory study. Before describing the routing procedure, we examine how changes to different attributes appear across VAR scales. Using GPT-Image-2, we construct two aligned image triplets: one varies the intensity of a woman's smile, and the other varies the density of fog in a street scene. Within each triplet, the remaining scene elements are kept as consistent as possible. We encode the images with Infinity's frozen VAE, decode representations accumulated up to selected scales, and compute spatial bit-disagreement maps, as shown in Figure 1.

Figure 1 provides two observations. First, the examples differ in both spatial extent and scale of onset. The smile triplet has identical codes at the initial 1×1 scale, with differences appearing mainly around the mouth at intermediate scales. In contrast, the fog triplet already differs at the 1×1 scale, and the differences extend across a broader part of the image. Second, both triplets exhibit dense bit differences at the finest scale. Thus, the Hamming distance at the final scale alone does not reveal when visible attribute changes first emerge. Together, these examples motivate considering spatial extent and generation scale jointly when routing an attribute edit.

4.1 Spatial Attribute Routing

The goal of spatial attribute routing is to determine where an attribute edit should take effect in the image. We represent this spatial extent with a soft mask $M \in [0, 1]$, where larger values indicate locations that should receive stronger editing. To infer this mask directly from the model, we keep the original generation trajectory fixed and compare the model responses under the base prompt P and the edited prompt P^a. Their prediction difference, denoted by Δr, indicates how introducing

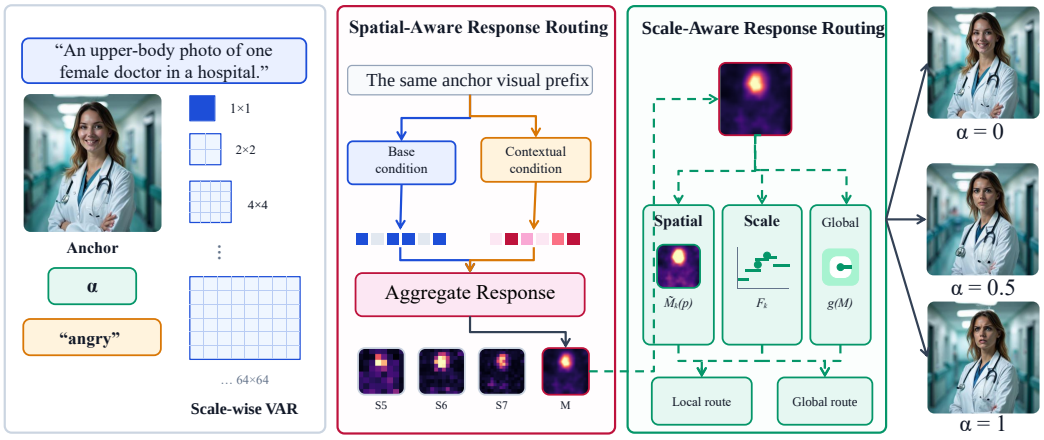


Figure 2: Overview of S2AR.

attribute a changes the model's response at each spatial location. However, not every large response difference corresponds to an actual edit: the edited prompt may simply make the model more confident in the bit value that is already present in the source image. For this reason, using the magnitude $|\Delta r|$ alone gives a noisy editing map, since it cannot distinguish changes that preserve the current image state from changes that actively push the representation toward a different state. We therefore retain only response changes that push against the observed source bit $z^0 \in \{0, 1\}$, i.e., changes that move the prediction toward a bit flip. Defining the source polarity as $s = 2z^0 - 1$, the resulting directional response is:

$$A = \mathrm{ReLU}\left(-s \cdot \Delta r\right). \tag{4}$$

This operation has a simple interpretation. If the source bit is 0, only changes that increase its prediction are retained; if the source bit is 1, only changes that decrease its prediction are retained. Thus, A measures where the edited prompt is actually trying to change the source representation rather than merely reinforcing it.

The directional response is still defined over individual bit channels and generation scales. To obtain a stable spatial estimate, we first average A across bit channels and then collect the responses from intermediate scales $\{5, 6, 7\}$, where attribute-related spatial structure is most clearly expressed. We resize these responses to a common spatial resolution and aggregate them into a single response map R, which describes where the target attribute tends to induce representation changes. Finally, we normalize R and suppress weak responses so that locations strongly associated with the attribute remain close to 1, while small background variations are mapped toward 0. The resulting mask $M \in [0, 1]$ therefore specifies where the attribute signal should be injected during generation.

4.2 Scale Attribute Routing

Spatial routing determines where to edit, but it does not determine at which generation scales the edit should be introduced. Scale attribute routing addresses this second question by assigning an editing weight β_k to each scale k. Our key observation is that the appropriate scale depends on the spatial extent of the attribute. A localized response, such as a change around the mouth, is poorly represented on a very coarse grid, whereas a scene-wide response, such as fog or illumination, can remain recognizable even after strong spatial downsampling. We therefore use the ability of scale k to preserve the spatial response map R as a measure of whether that scale is suitable for editing the attribute. Concretely, we downsample R to the spatial resolution (h_k, w_k) of scale k and then upsample it back to the original resolution, producing a reconstructed response map $\hat{R}_k$. If $\hat{R}_k$ remains close to R, then the spatial structure of the attribute can already be represented at scale k. If the reconstruction changes substantially, the scale is too coarse to faithfully express that attribute. We convert this reconstruction fidelity into the scale weight β_k:

$$\beta_k = \begin{cases} 0, & k = 1, \\ \max\left(0, 1 - \frac{\|R - \hat{R}_k\|_1}{\|R\|_1 + \epsilon}\right), & k > 1. \end{cases} \tag{5}$$

Accordingly, β_k is large when scale k preserves the response pattern well and becomes small when downsampling destroys important spatial structure. We additionally set $\beta_1 = 0$ so that attribute editing does not modify the initial 1×1 representation, which primarily determines the overall image composition.

The scale weights above determine when the edit should be injected, but we also need to distinguish edits that should remain spatially localized from those that should modify the scene as a whole. We therefore compute a globalness score $g \in [0, 1]$ directly from the spatial response R:

$$g = \frac{(\sum_u R(u))^2}{N \sum_u R(u)^2 + \epsilon}, \tag{6}$$

where N is the number of spatial positions. The score is small when the response is concentrated in a limited region and becomes large when the response is distributed broadly across the image. Thus, a localized attribute such as a smile produces a small g, whereas a scene-wide attribute such as fog or lighting produces a larger g.

4.3 Generation Pipeline

Figure 2 summarizes the generation pipeline. Given the spatial mask M, the scale weights $\{\beta_k\}$, and the globalness score g, we route the attribute signal through two complementary pathways. The first pathway modifies spatially localized details, while the second adjusts scene-level properties when the target attribute requires a global change. Both pathways are controlled by the shared scale weight β_k, ensuring that the attribute is introduced only at scales capable of representing it reliably.

Local detail injection. At scale k, let $\Delta\mathbf{C}_k$ denote the shift in the cross-attention output induced by replacing the base prompt with the edited prompt. Rather than applying this update uniformly, we inject it selectively where the spatial mask indicates that the target attribute should appear:

$$\tilde{\mathbf{C}}_k = \mathbf{C}_k^0 + \alpha\beta_k M_k \cdot \Delta\mathbf{C}_k, \tag{7}$$

where M_k is the spatial mask M resized to the resolution of scale k, and $\alpha \in [0, 1]$ controls the overall editing strength.

Global scene modulation. The globalness score g gates this pathway: for localized attributes, $g \approx 0$, which preserves the original global conditioning and leaves the background unaffected; for scene-level attributes, a larger g allows the edited prompt to modify global properties such as scene tone and illumination. At scale k, we interpolate between the base and edited prompt embeddings using the weight $\lambda_k = \alpha\beta_k g$.

5 Experiments

5.1 Experimental Setup

CACBench. To evaluate continuous attribute control across diverse scenarios, we construct CACBench, which comprises 21 attributes. The benchmark contains three subsets: facial expressions (6 attributes), object properties (6 attributes), and scene conditions (9 attributes). For each attribute, we evaluate eight base prompts across five control intensities ($\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$) under five fixed random seeds. All evaluated methods share the same base prompts, attribute keywords, control intensities, and random seeds. Conceptually, these tasks cover both local and global modifications: local edits are divided into non-rigid deformations (such as facial expressions) and rigid transformations (such as objects), whereas scenes correspond to global alterations.

Baselines. Using the same Infinity-2B backbone Han et al. (2025) and identical generative anchor points, we compare our method with visual autoregressive editing baselines, namely AREdit Wang et al. (2025) and BitResEdit (Zhang et al., 2026), both of which operate directly on generated anchor images. In addition, we train a LoRA baseline Gandikota et al. (2024); Hu et al. (2021) on Infinity-2B following the design of Concept Sliders. All experiments are conducted on NVIDIA A100 40 GB GPUs.



Figure 3: Qualitative comparison across three CACBench control settings.

Table 1: Quantitative comparison across the three control tasks on CACBench (best in **bold**, second-best underlined).

Task	Method	ΔCLIP↑	$\rho\uparrow$	Reversal↓	DINO↑	LPIPS↓
Expressions	Concept Sliders	0.043	0.889	0.154	0.689	0.545
Expressions	AREdit	0.020	0.461	0.371	0.273	0.365
Expressions	BitResEdit	**0.055**	0.892	0.182	0.726	0.469
Expressions	Ours	0.048	**0.908**	**0.142**	**0.941**	**0.238**
Object	Concept Sliders	0.034	0.858	0.189	0.526	0.642
Object	AREdit	0.013	0.655	0.277	0.706	0.236
Object	BitResEdit	**0.041**	0.873	0.173	0.622	0.564
Object	Ours	0.037	**0.888**	**0.158**	**0.824**	**0.218**
Scene	Concept Sliders	0.065	0.801	0.204	0.440	0.686
Scene	AREdit	0.017	0.458	0.364	0.592	0.314
Scene	BitResEdit	0.066	0.875	0.151	0.572	0.579
Scene	Ours	**0.068**	**0.890**	**0.139**	**0.765**	**0.286**

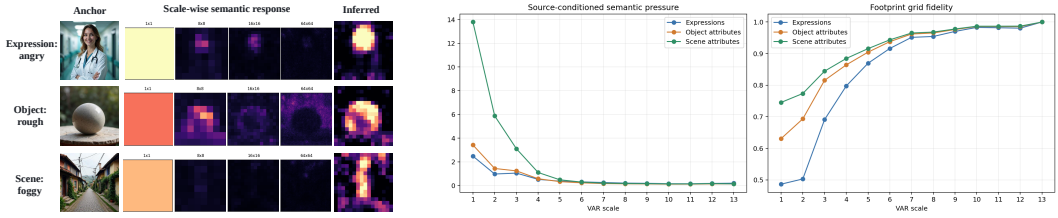
Evaluation Metrics. We assess continuous attribute control across three dimensions: semantic fidelity, monotonic progression, and identity preservation. To quantify semantic fidelity, we calculate ΔCLIP, defined as the change in the CLIP (Radford et al., 2021) score margin between the target and source text prompts from the unedited base ($\alpha = 0$) to the maximum control strength ($\alpha = 1$). To determine whether attribute changes remain strictly monotonic, we compute the Spearman rank correlation coefficient ρ between the control parameter α and the CLIP margin, alongside the reversal rate, which measures the frequency of margin decreases between consecutive steps. Finally, we measure structure and identity preservation using DINOv2 (Oquab et al., 2023) feature similarity and LPIPS (Zhang et al., 2018) between the fully edited image ($\alpha = 1$) and the original reference ($\alpha = 0$).

5.2 Main Results

Figure 3 compares qualitative results across the three CACBench tasks. For facial expression editing, our method smoothly scales emotional intensity while preserving identity, facial geometry, and attire. In contrast, the three baselines exhibit noticeable identity drift: BitResEdit and the LoRA baseline visibly distort facial features, whereas AREdit produces minimal emotional change. On object property editing, our approach accurately preserves the original padlock geometry, whereas AREdit distorts its shape, the LoRA baseline noticeably enlarges it, and BitResEdit alters both its size and orientation. For scene condition editing, the LoRA baseline fails by generating overexposed images, while AREdit yields little visible change. Although BitResEdit retains more scene structure

Table 2: Ablation study of key components in S2AR. The best results are highlighted in **bold**, and the second-best results are underlined.

Configuration	ΔCLIP↑	ρ ↑	Reversal↓	DINO↑	LPIPS↓
Full method	**0.045**	**0.735**	**0.215**	**0.945**	**0.105**
w/o Spatial Response	0.042	0.722	0.225	0.718	0.397
w/o Scale Response	0.010	0.466	0.365	0.939	0.114
Unlocked 1×1 root	0.038	0.668	0.269	0.764	0.368
Isolated attribute condition	0.013	0.371	0.390	0.878	0.256



(a) Target region identification

(b) Scale-wise signal profiles

Figure 4: Analysis of spatial target identification and scale-wise dynamics. (a) Response maps across representative VAR scales progressively isolate target regions for expression, object, and scene attributes. (b) Signal dynamics across 13 VAR scales: semantic pressure (left) and grid fidelity (right) reveal that global scene attributes activate early coarse scales, whereas localized expressions concentrate at finer scales.

than the other baselines, it undesirably alters the road surface and roadside environment. Across increasing editing strengths, our method produces consistent, progressive edits while keeping the scene layout intact.

Table 1 reports quantitative evaluations across the three CACBench tasks. Our method achieves the best overall balance among attribute steering, content preservation, and continuous control. Specifically, it consistently maintains structure and identity across tasks (achieving a DINO similarity of 0.941 and an LPIPS of 0.238 on facial expressions) while ensuring monotonic steering, as evidenced by the lowest reversal rates and highest rank correlation ρ. Although BitResEdit modifies facial and object attributes to a greater degree, it substantially degrades global scene layout and background fidelity (reducing scene DINO to 0.572 and increasing LPIPS beyond 0.46), while frequently reversing control directions. Our method, by contrast, achieves the highest text-image alignment on the scene benchmark (ΔCLIP of 0.068) without compromising underlying structures. The other baselines exhibit clear limitations: Concept Sliders introduces heavy visual artifacts and structural breakdown (0.440 DINO and 0.686 LPIPS on scenes), whereas the low LPIPS of AREdit reflects an inability to produce noticeable edits. Overall, these results demonstrate that our training-free method enables smooth and monotonic attribute steering with high fidelity, without relying on external masks.

5.3 Mechanism Analysis

Ablation of Key Components. Table 2 validates the contribution of each design choice. Disabling spatial response routing significantly degrades background preservation (DINO drops from 0.945 to 0.718, and LPIPS increases from 0.105 to 0.397), confirming that spatial localization is essential for protecting unaffected regions. Omitting scale routing substantially impairs steering performance (ΔCLIP drops from 0.045 to 0.010, and monotonicity falls to $\rho = 0.466$), indicating that uniform scale application fails to accommodate diverse feature hierarchies. Additionally, unlocking the 1×1 root bit causes early trajectory drift (DINO drops to 0.764), while replacing contextual conditions with isolated tokens induces conditioning distribution shifts (ΔCLIP drops to 0.013). These results show that all routing components are necessary to balance controllability and identity preservation.

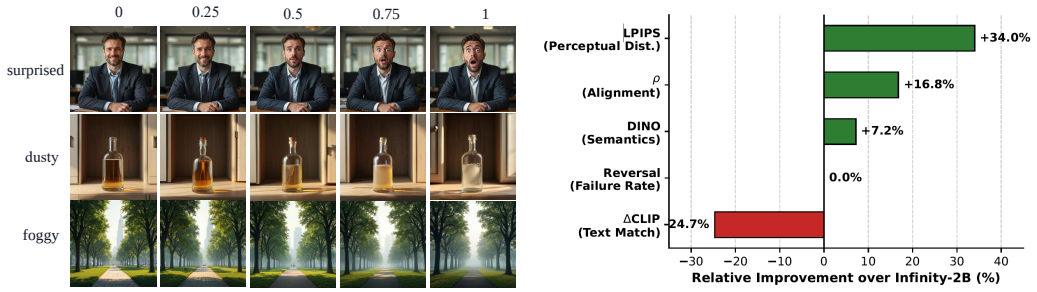


Figure 5: Generalization to Infinity-8B.

Table 3: Efficiency comparison on an NVIDIA A100 40GB GPU.

Method	Base Model	Latency (s)	Peak Memory (GiB)
Concept Sliders	2B	9.79	15.76
AREdit	2B	10.45	15.85
BitResEdit	2B	12.62	20.01
Ours	2B	**8.83**	**13.03**

Spatial and Scale Routing Dynamics. To understand how our routing adapts across heterogeneous attributes, we analyze its spatial and scale-wise dynamics in Figure 4. As shown in Figure 4(a), the inferred response maps progressively localize attribute-relevant regions across VAR scales without requiring manual masks. Figure 4(b) further reveals distinct scale-wise behaviors across attribute types: global scene attributes exhibit stronger semantic influence at earlier, coarser scales, whereas localized facial expressions concentrate more strongly at later, finer scales, where grid fidelity is higher. These results show that our routing mechanism adapts both the spatial extent and scale allocation of the editing signal according to the semantics of different attributes.

Transferability to Infinity-8B. To evaluate the generalizability of S2AR, we apply it directly to the larger Infinity-8B model without additional tuning. As shown in Fig. 5 (left), the controller transfers effectively to the 8B backbone, producing smooth and continuous edits across diverse semantic directions. Quantitatively (Fig. 5, right), the 8B model yields higher control monotonicity (ρ increases by 16.8%) and structural consistency (DINO increases by 7.2%), while reducing perceptual distortion by 34.0%. Although ΔCLIP drops by 24.7%, we attribute this to the 8B model suppressing over-editing. Overall, our method achieves a better trade-off between attribute controllability and content preservation at a larger scale.

Inference Latency and Memory Footprint. Table 3 reports the operational efficiency on a single NVIDIA A100 (40GB) GPU. Because our method requires neither offline fine-tuning nor test-time optimization, it completes a 5-level editing trajectory in 8.83 seconds with a peak memory footprint of 13.03 GiB on the 2B backbone. In comparison, the LoRA-based Concept Sliders takes 9.79 seconds and consumes 15.76 GiB while incurring additional training overhead. Our method also requires substantially less memory than AREdit (15.85 GiB) and BitResEdit (20.01 GiB).

6 Conclusion

In this paper, we investigated continuous visual attribute control in visual autoregressive models. We observed that early semantic perturbations propagate across scales, leading to identity and compositional drift. To address this issue, we introduced S2AR, a training-free framework for continuous attribute control. Specifically, S2AR constructs source-conditioned spatial localization maps to localize attribute-relevant regions and incorporates a spatial-scale routing mechanism to restrict edits to the appropriate hierarchical scales. By injecting scaled prediction residuals into the generative process, our method achieves smooth, continuous control without requiring parameter tuning or external models. Experiments on CACBench show that S2AR provides high controllability while preserving subject identity, composition, and background.

AI USE STATEMENT

Generative AI tools were used for language polishing, typo correction, and generating synthetic images for exploratory analysis. The authors reviewed all AI-assisted content and take responsibility for the final manuscript and research findings.

REFERENCES

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13231–13241. IEEE, 2025.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

Die Chen, Zhongjie Duan, Zhiwen Li, Cen Chen, Daoyuan Chen, Yaliang Li, and Yinda Chen. Attrictrl: A generalizable framework for controlling semantic attribute intensity in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Pin-Yen Chiu, I Fang, Jun-Cheng Chen, et al. Text slider: Efficient and plug-and-play continuous concept control for image/video synthesis via lora adapters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 613–622, 2026.

Rotem Ezra, Hedi Zisling, Nimrod Berman, Ilan Naiman, Alexey Gorkor, Liran Nochumsohn, Eliya Nachmani, and Omri Azencot. Freesliders: Training-free, modality-agnostic concept sliders for fine-grained diffusion control in images, audio, and video. *arXiv preprint arXiv:2511.00103*, 2025.

Hongyi Fang, Chuwen Xie, Benjia Zhou, Yu-Xuan Qiu, Chenggong Hu, Zhibin Wang, Chao Chen, Jianbin Qin, and Rui Mao. Model the edit, not the image: Visual autoregressive editing from a source-centric perspective. *arXiv preprint arXiv:2608.09057*, 2026.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15994–16003. IEEE, 2025.

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *2024 ieee 18th international conference on automatic face and gesture recognition (fg)*, pp. 1–9. IEEE, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity∞: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744. IEEE, 2025.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Dongyang Jin, Ryan Xu, Jianhao Zeng, Rui Lan, Yancheng Bai, Lei Sun, and Xiangxiang Chu. Semantic context matters: Improving conditioning for autoregressive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30402–30413, 2026.

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

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 8450–8474, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/185087ea328b4f03ea8fd0c8aa96f747-Paper-Conference.pdf.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 289–299. IEEE, 2023.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. In *International Conference on Learning Representations*, volume 2026, pp. 85702–85723, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *European Conference on Computer Vision*, pp. 469–487. Springer, 2024.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9240–9249. IEEE, 2020.

Youngwoo Shin, Junmo Kim, et al. Ssg: Scaled spatial guidance for multi-scale visual autoregressive generation. In *International Conference on Learning Representations*, volume 2026, pp. 26774–26799, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Deepak Sridhar and Nuno Vasconcelos. Prompt sliders for fine-grained control, editing and erasing of concepts in diffusion models. In *European Conference on Computer Vision*, pp. 17–31. Springer, 2024.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer, 2024. *URL https://arxiv. org/abs/2410.10812*.

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

Yufei Wang, Lanqing Guo, Zhihao Li, Jiaxing Huang, Pichao Wang, Bihan Wen, and Jian Wang. Training-free text-guided image editing with visual autoregressive model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17577–17586. IEEE, 2025.

Tao Xia, Jiawei Liu, Yukun Zhang, Ting Liu, Wei Wang, and Lei Zhang. Rethinking structure preservation in text-guided image editing with visual autoregressive models. *arXiv preprint arXiv:2603.28367*, 2026.

Dongli Xu, Aleksei Tiulpin, and Matthew Blaschko. Softcfg: Uncertainty-guided stable guidance for visual autoregressive model. In *International Conference on Learning Representations*, volume 2026, pp. 13869–13887, 2026a.

Ryan Xu, Dongyang Jin, Yancheng Bai, Rui Lan, Xu Duan, Lei Sun, and Xiangxiang Chu. Scalar: Scale-wise controllable visual autoregressive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11379–11387, 2026b.

Xindi Yang, Yicheng Wu, Cheng Zhang, Jianfei Cai, and Tien-Tsin Wong. Attribute token arithmetic: Disentangled and continuous semantic control for visual autoregressive models. In *European Conference on Computer Vision*, pp. 297–314. Springer, 2026.

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

Weixin Ye, Hongguang Zhu, Wei Wang, Yahui Liu, Mengyu Wang, and Xuecheng Nie. All-in-one slider for attribute manipulation in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18693–18702, 2026.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Arman Zarei, Samyadeep Basu, Mobina Pournemat, Sayan Nag, Ryan A Rossi, and Soheil Feizi. Slideredit: Continuous image editing with fine-grained instruction control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14430–14439, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shengqiang Zhang, Ruotong Liao, Volker Tresp, Barbara Plank, and Hinrich Schütze. Edit the bits, diff the codes: Bitwise residual editing for visual autoregressive models. *arXiv preprint arXiv:2606.13558*, 2026.

Zixin Zhu, Kevin Duarte, Mamshad Nayeem Rizve, Chengyuan Xu, Ratheesh Kalarot, and Junsong Yuan. Compslider: Compositional slider for disentangled multiple-attribute image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.