VISTA: TEST-TIME COMPOSITIONAL ALIGNMENT FOR VISUAL AUTOREGRESSIVE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual autoregressive (VAR) models offer fast, high-quality text-to-image generation but still fail at attribute binding and spatial relations. Existing VAR test-time methods search sampled trajectories or apply generic logit guidance but do not directly optimize compositional constraints; gradient-based diffusion methods do not transfer to stateful, discrete, multi-resolution VAR sampling. We introduce **VISTA** (**Vi**sual Autoregressive **S**emantic **T**est-time **A**lignment), the first gradient-based test-time framework for compositional alignment in next-scale autoregressive image generation. Built on Infinity, VISTA optimizes intermediate representations through the frozen transformer without parameter updates or training. It stabilizes optimization across scales and exposes a common interface into which any differentiable constraint on cross-attention can be plugged. Across two benchmarks and model scales, VISTA improves every targeted category, raising the mean by nearly 20% on a 2B backbone and almost 6% on an 8B backbone, especially on spatial relations. In human evaluation, VISTA wins 83% of decisive compositionality comparisons by plurality vote, with no significant preference in perceived quality; independently, ImageReward scores its outputs 20% higher. VISTA enables the 2B model to surpass a backbone four times its size, showing much of the compositional gap is recoverable at test time.

1 INTRODUCTION

Latent diffusion models Rombach et al. (2022) and their transformer-based successors remain the dominant approach to open-domain text-to-image synthesis. Visual autoregressive (VAR) models offer a recent alternative: VAR Tian et al. (2024) recasts autoregressive image generation as coarse-to-fine next-scale prediction rather than raster-scan token prediction Esser et al. (2021), producing increasingly resolved token maps conditioned on all coarser ones. This formulation allowed transformer-based Vaswani et al. (2017) autoregressive models to surpass diffusion transformers in quality, speed, and scalability. Infinity Han et al. (2025a) extends it with a bitwise, infinite-vocabulary tokenizer and self-correction, reaching quality competitive with leading diffusion models.

Despite this progress, high visual quality does not guarantee compositional reliability. A systematic comparison of six text-to-image systems finds that next-scale VAR models such as Infinity fail on attribute binding, spatial reasoning, and relational tasks, and that these failures persist with model scale Shahabadi et al. (2025). Visual fidelity and parameter count therefore do not imply compositional understanding. Because retraining a backbone is expensive and model-specific, a growing body of work instead intervenes at inference time.

In diffusion models, inference-time compositional alignment is already well established. Building on evidence that cross-attention encodes correspondence between image regions and text Hertz et al. (2022), Attend-and-Excite Chefer et al. (2023) and SynGen Rassin et al. (2023) optimize the diffusion latent during sampling to reshape attention toward prompt constraints. Subsequent work refines these objectives for attribute binding Meral et al. (2024); Li et al. (2023), concept separation Agarwal et al. (2023), and spatial relations Rezaei et al. (2025); Han et al. (2025b). These methods share the premise that a frozen model's cross-attention provides a differentiable handle on compositional structure, making inference-time steering cheaper and more portable than retraining.

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.

Test-time intervention has also reached autoregressive image generation through search and logit guidance. TTS-VAR Chen et al. (2025) searches next-scale trajectories using coarse-scale diversity and fine-scale reward selection; GridAR Park et al. (2026) progressively prunes partial raster-scan generations; and ScalingAR Haodong Chen et al. (2025) adapts test-time scaling to next-token generation. For next-scale VARs, SSG Shin et al. (2026) amplifies scale-specific residual detail, while VPG Liao et al. (2026) contrasts predictions from genuine and corrupted visual prefixes. These guidance methods target generic scale refinement or prefix consistency rather than explicit compositional constraints. Selection is most effective when satisfactory candidates already occur with appreciable probability, but offers limited help when the backbone fails systematically. Gradient steering can directly modify the trajectory, but cannot be applied unchanged to next-scale VAR generation. Infinity produces discrete tokens over increasing resolutions and maintains persistent state on which later scales depend. Consequently, global gradient normalization dilutes updates across increasingly large token maps until individual elements no longer affect sampling, while repeated optimization passes alter the attention cache read by subsequent scales.

We introduce **VISTA** (**Vi**sual Autoregressive **S**emantic **T**est-time **A**lignment), the first gradient-based test-time framework for compositional alignment in next-scale VAR generation. Applied to Infinity, VISTA optimizes intermediate token representations through the frozen transformer using differentiable cross-attention capture and cache-safe, scale-aware optimization. The same capture and optimization machinery serves each instantiated objective. A pluggable registry supports differentiable cross-attention constraints; we instantiate it for attribute binding, planar relations, and depth ordering.

On T2I-CompBench, VISTA improves every targeted category, including an 89% relative gain on 2D spatial relations and 11–21% gains on color, texture, and shape binding. On GenEval, VISTA raises Position from 0.250 to 0.745, whereas TTS-VAR reports a 7.5% Position gain on Infinity-2B under its own protocol Chen et al. (2025). VISTA on Infinity-2B surpasses the $4\times$ larger Infinity-8B on four of six targeted categories and on their average, while applying VISTA to Infinity-8B improves that model further. Human evaluation prefers VISTA in 83% of decisive compositionality comparisons and finds no significant preference in perceived quality; independently, ImageReward scores its outputs nearly 20% higher. These gains trade against inference cost, which VISTA exposes through two explicit controls, while requiring no additional training, trainable parameters, reward model, or task-specific data. Appendix A discusses related work in full.

Our contributions are threefold. (i) Gradient-based alignment for next-scale VARs. We introduce the first gradient-based test-time framework for compositional alignment in next-scale VARs, requiring no additional training, parameters, reward model, or candidate search. **(ii) Stable optimization inside stateful generation.** We identify the failure modes that make latent optimization unstable in multi-scale VAR sampling and introduce shared capture and optimization mechanisms that address them (Secs. 3.2 and 3.4). **(iii) A reusable objective interface.** We make differentiable cross-attention objectives developed for diffusion applicable to next-scale VARs, opening a new direction for test-time objective design. We demonstrate this interface using prior binding and planar objectives together with a new boundary-containment depth constraint.

2 PRELIMINARIES

VARs generate coarse-to-fine. VAR Tian et al. (2024) represents an image as a sequence of discrete token maps $r_1, \ldots, r_S$ at increasing resolutions $h_s \times w_s$. Each r_s indexes a codebook V learned by a VQ tokenizer, and generation factorizes autoregressively over scales:

$$p(r_1, \ldots, r_S \mid c) = \prod_{s=1}^{S} p(r_s \mid r_1, \ldots, r_{s-1}, c). \tag{1}$$

Infinity Han et al. (2025a) extends VAR from class conditioning to free-form text by taking c from a frozen text encoder, and replaces the finite codebook with a bitwise, infinite-vocabulary tokenizer whose labels are predicted one bit at a time. Its default 13-scale schedule progresses from 1×1 and 2×2 grids through 4×4, 6×6, and 8×8, ultimately reaching 64×64.

VISTA intervenes before irreversible sampling in a stateful process. At scale s, Infinity upsamples the previously sampled and quantized codes $r_{<s}$ to resolution $h_s \times w_s$, producing a continuous

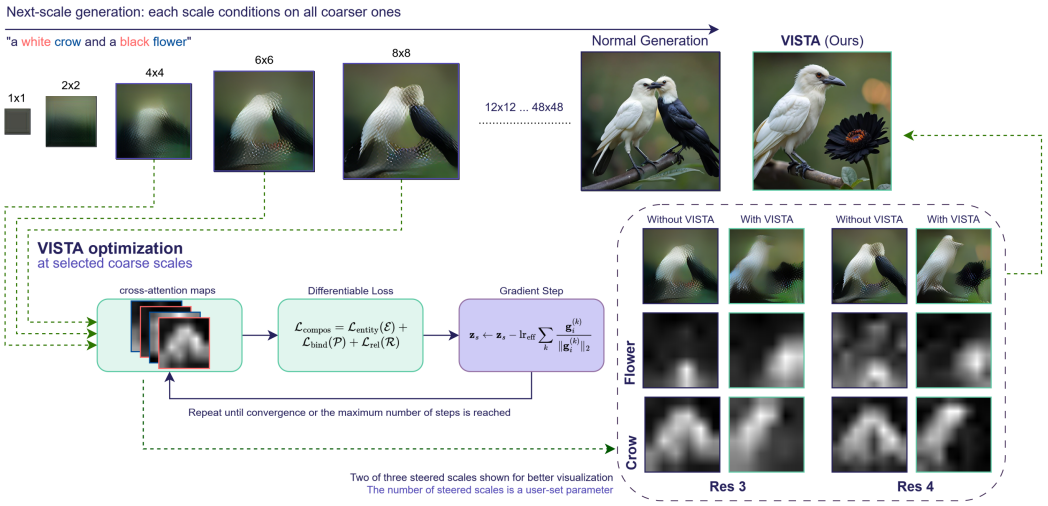

Figure 1: VISTA on "a white crow and a black flower." Generation proceeds coarse to fine, with every scale conditioned on all coarser ones. At the outlined scales, VISTA evaluates differentiable constraints on the model's cross-attention and updates $\mathbf{z}_s$ until all active gradients fall below their thresholds or the step budget is exhausted. **Right:** *flower* and *crow* attention at the 6×6 and 8×8 scales. Without steering, the maps remain collapsed and a second bird replaces the flower; with steering, they separate and remain distinct through the unsteered finer scales. $\mathcal{L}_{\text{entity}}$ is shown for completeness but is not instantiated in our experiments.

representation $\mathbf{z}_s \in \mathbb{R}^{L_s \times C}$, where $L_s = h_s w_s$. A transformer processes $\mathbf{z}_s$ through self-attention over spatial positions and cross-attention to c, then produces logits from which r_s is sampled using top-k/top-p sampling. The resulting tokens are quantized and passed to scale $s+1$. Keys and values computed at each scale are cached and read by every later scale, so generation carries state across the entire schedule rather than restarting at each resolution. Consequently, $\mathbf{z}_s$ is a transient, per-scale representation rather than a single latent refined throughout generation as in diffusion; intervention occurs before an irreversible discrete sampling step, and every optimization forward pass can modify state needed by later scales. These properties motivate the cache-safe optimization of Secs. 3.2–3.4.

Compositionality is constraint satisfaction. We treat a prompt as asserting constraints over its entities: each entity's *cardinality*, its bound *attributes*, and its *relations* to other entities. A generation fails compositionally when it violates one or more of these constraints, independently of its overall visual quality. Such failures remain a central challenge even for strong generative models: high visual fidelity does not ensure that all requested entities, attributes, and relations are jointly satisfied Huang et al. (2023); Ghosh et al. (2023); Shahabadi et al. (2025).

3 Method

3.1 Overview

VISTA separates reusable optimization machinery from pluggable compositional objectives. Cross-attention provides the control signal: because image tokens query the text conditioning $\mathbf{c}$, its maps expose which text tokens influence each spatial region Hertz et al. (2022). At selected scales, VISTA treats Infinity's continuous input representation $\mathbf{z}_s$ as an optimization variable and steers it until these maps better satisfy constraints extracted from the prompt. The backbone and text encoder remain frozen, and no training data is required. The optimized representation $\widetilde{\mathbf{z}}_s$ replaces $\mathbf{z}_s$ before discrete sampling; the resulting token map r_s then conditions every finer scale. VISTA first captures cross-attention differentiably (Sec. 3.2), expresses prompt constraints as objectives over it (Sec. 3.3), and then optimizes those objectives safely inside stateful generation (Sec. 3.4). Figure 1 summarizes the method.

3.2 Differentiable Cross-Attention Capture

Cross-attention becomes a differentiable control interface. Infinity normally evaluates cross-attention inside a frozen, cached sampling loop that retains no gradient path to $\mathbf{z}_s$. We reconstruct the attention probabilities during an optimization forward pass. For cross-attention layer ℓ and head h, let $\mathbf{h}_\ell^{(s)}(\widetilde{\mathbf{z}}_s)$ denote the current layer input induced by the representation being optimized. We compute

$$\mathbf{q}_{\ell h} = W_{\ell h}^{Q}\mathbf{h}_\ell^{(s)}(\widetilde{\mathbf{z}}_s), \qquad \mathbf{k}_{\ell h} = \mathrm{sg}\big(W_{\ell h}^{K}\mathbf{c}\big)\,, \qquad \mathbf{A}_{\ell h}^{(s)} = \mathrm{softmax}\left(\frac{\mathbf{q}_{\ell h}\mathbf{k}_{\ell h}^{\top}}{\sqrt{d_h}}\right), \tag{2}$$

where d_h is the head dimension and $\mathrm{sg}(\cdot)$ denotes stop-gradient. Gradients flow through each layer's query and hidden representation to $\widetilde{\mathbf{z}}_s$, but not into the fixed text stream or model parameters. Averaging $\mathbf{A}_{\ell h}^{(s)}$ over all cross-attention layers and heads gives $\mathbf{A}^{(s)} \in \mathbb{R}^{L_s \times L_k}$, where L_k is the text-token length. Together with the losses defined in Sec. 3.3, the captured cross-attention maps provide the signal VISTA uses to improve alignment between the text prompt and the generated image.

3.3 Compositional Objectives

Prompt constraints become attention objectives. A prompt is compositionally satisfied only when every asserted constraint holds: each entity appears in the stated number, each attribute binds to its intended entity, and each relation holds between the correct pair. Cross-attention connects these constraints to image layout: column t of $\mathbf{A}^{(s)}$ scores how strongly each spatial position attends to text token t. A constraint on an entity's placement can therefore be written as a differentiable function of its attention mass.

A common interface separates objectives from optimization. The parser returns explicitly counted entities $\mathcal{E}$, attribute–noun pairs $\mathcal{P}$, and ordered relation triples $r = (X, \rho, Y) \in \mathcal{R}$. The relation phrase ρ assigns each triple to one of three disjoint sets: $\mathcal{R}_{\mathrm{2D}}$ for image-plane relations, $\mathcal{R}_{\mathrm{dep}}$ for viewing-axis relations, and $\mathcal{R}_{\mathrm{ns}}$ for non-spatial relations. The general objective is

$$\mathcal{L}_{\mathrm{compos}} = \mathcal{L}_{\mathrm{entity}}(\mathcal{E}) + \mathcal{L}_{\mathrm{bind}}(\mathcal{P}) + \mathcal{L}_{\mathrm{rel}}(\mathcal{R}), \qquad \mathcal{L}_{\mathrm{rel}} = \underbrace{\mathcal{L}_{\mathrm{2D}} + \mathcal{L}_{\mathrm{depth}}}_{\mathcal{L}_{\mathrm{spatial}}} + \mathcal{L}_{\mathrm{ns}}, \tag{3}$$

An empty set contributes zero. Our reported configuration instantiates $\mathcal{L}_{\mathrm{bind}}$, $\mathcal{L}_{\mathrm{2D}}$, and $\mathcal{L}_{\mathrm{depth}}$; $\mathcal{L}_{\mathrm{entity}}$ and $\mathcal{L}_{\mathrm{ns}}$ remain inactive.

For a text-token group T, define its attention map at scale s as

$$\mathbf{m}_T^{(s)} = \frac{1}{|T|}\sum_{t \in T} \mathbf{A}_{:,t}^{(s)} \in \mathbb{R}^{L_s}. \tag{4}$$

We drop the scale superscript when evaluating one scale and write $\mathbf{m}_X$ for entity X. The interface accepts any differentiable scalar objective over these maps.

Existing diffusion objectives transfer through the interface. The interface accepts any differentiable scalar objective over the captured attention maps without changing the capture or optimization machinery. Throughout this work, we instantiate it with objectives for attribute binding, planar relations, and depth ordering. For binding, $\mathcal{L}_{\mathrm{bind}}$ adapts CONFORM's multi-positive contrastive objective Meral et al. (2024): an attribute and its noun form a positive pair, while other pairs provide negatives, pulling each attribute toward its intended entity and away from competing objects. For planar relations, $\mathcal{L}_{\mathrm{2D}}$ adopts PSG's probability-of-superiority objective Rezaei et al. (2025), which interprets two attention maps as spatial distributions and drives their ordering along the asserted axis. We introduce the occlusion-based depth objective below. Appendix B gives the complete definitions and provenance of the binding and planar terms, while Appendix E evaluates alternative objectives under the same framework.

Depth ordering is induced through occlusion geometry. Cross-attention has no explicit depth coordinate, but front–back order leaves an asymmetric footprint in the image plane. For "B behind A," our depth objective combines a separation term that prevents the two attention supports from

collapsing and a boundary-containment term that encourages the portion of B facing A to pass through A's support:

$$\mathcal{L}_{\text{depth}} = \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{bc}}\mathcal{L}_{\text{bc}}. \tag{5}$$

This does not estimate metric depth; it promotes the overlap and foreground–background asymmetry through which the requested order becomes visible. Appendix C gives the complete construction, motivation, and hyperparameters.

This objective is one instantiation proposed alongside VISTA, not a requirement of the framework. Appendix E holds the capture and optimization machinery fixed and compares five depth objectives; every alternative enters through the same interface, and boundary containment performs best in that diagnostic comparison.

3.4 Stable Optimization in a Stateful Generation Process

Given the active objectives above, VISTA must update $\mathbf{z}_s$ without corrupting Infinity's persistent state, suppressing useful gradients, or spending equal computation on every scale.

Cache isolation preserves the sampling trajectory. Each auxiliary optimization forward writes to the self-attention cache used by later scales. VISTA therefore snapshots the cache before optimization, restores it before every auxiliary forward, and restores it once more before ordinary sampling resumes. The backbone remains frozen, and the final sampling forward sees the same pre-scale state it would have received without optimization.

Per-token directions balance heterogeneous objectives. Let t index optimization steps, $i \in \{1, \dots, L_s\}$ spatial tokens, and k active objectives. For objective $\mathcal{L}_k$, define

$$\mathbf{g}_i^{(k,t)} = \nabla_{\widetilde{\mathbf{z}}_{s,i}^{(t)}}\mathcal{L}_k, \qquad \overline{g}^{(k,t)} = \frac{1}{L_s}\sum_{i=1}^{L_s}\left\|\mathbf{g}_i^{(k,t)}\right\|_2. \tag{6}$$

Let $\mathcal{K}_t = \{k : \overline{g}^{(k,t)} \geq \tau_k\}$, where τ_k is the activity threshold for objective k. Starting from $\widetilde{\mathbf{z}}_s^{(0)} = \mathbf{z}_s$, VISTA combines normalized gradient directions rather than incomparable loss values:

$$\mathbf{d}_i^{(t)} = \sum_{k\in\mathcal{K}_t}\frac{\mathbf{g}_i^{(k,t)}}{\left\|\mathbf{g}_i^{(k,t)}\right\|_2 + \epsilon}, \qquad \widetilde{\mathbf{z}}_{s,i}^{(t+1)} = \widetilde{\mathbf{z}}_{s,i}^{(t)} - \eta_s\mathbf{d}_i^{(t)}, \tag{7}$$

where $\epsilon > 0$ is a small constant. All objectives share one forward graph and contribute to one joint update.

The scale-aware step size is

$$\eta_s = \text{lr}_{\text{base}}\ \text{RMS}(\mathbf{z}_s)\sqrt{C}, \tag{8}$$

where C is the feature dimension and RMS is computed over the L_sC entries of $\mathbf{z}_s$. Per-token normalization prevents updates from vanishing as the grid grows, while RMS scaling adapts one base step size to different scales and prompts.

Compute follows violation at structurally important scales. The initial ratios $\overline{g}^{(k,0)}/\tau_k$ determine an effective ceiling $1 \leq N_{\text{eff}} \leq N_{\max}$: nearly satisfied scales receive few steps, while strongly violated ones receive the full budget. Optimization terminates earlier whenever $\mathcal{K}_t$ becomes empty. Appendix D gives the exact scheduler, and Sec. 4 reports the backbone-specific thresholds used.

VISTA additionally restricts intervention to scales where global structure is still forming. Perturbation analyses show that changing coarse VAR scales alters scene structure, whereas changing fine scales mainly affects local detail Tang et al. (2025). We therefore begin at 4×4, exclude the geometrically degenerate 1×1 and 2×2 grids, and steer the earliest three non-degenerate scales by default. Sec. 4.5 shows that gains saturate beyond this point.

Table 1: T2I-CompBench. VISTA and its Infinity baselines use four seeds; BoN and TTS-VAR are single runs. Prior-system rows are from Shahabadi et al. (2025). Every Infinity row uses ScaleKV (Sec. 4.1). VISTA steers three scales; TTS-VAR uses the authors' implementation at $N{=}1$ and $N{=}2$, and BoN uses $N{=}8$. Numeracy and Non-Spatial (shaded) are untargeted and unchanged by construction. Avg.$_6$ covers the six targeted categories and Avg.$_8$ all eight. **Bold**: best within each backbone group; underlined: best overall.

Model	Color	Texture	Shape	2D	3D	Cmplx	Num.	Non-S.	Avg.$_6$	Avg.$_8$
SDXL	0.593	0.519	0.466	0.215	0.341	0.319	0.504	0.311	0.409	0.409
PixArt-α	0.407	0.444	0.367	0.202	0.350	0.324	0.506	0.308	0.349	0.363
Flux-Dev	0.746	0.644	0.482	0.273	0.393	0.363	0.613	0.309	0.483	0.478
Flux-Schnell	0.725	0.683	0.559	0.271	0.373	0.364	0.604	0.312	0.496	0.486
Infinity-2B	0.749	0.632	0.475	0.234	0.405	0.384	0.572	0.310	0.480	0.470
+ BoN ($N{=}8$)	0.739	0.614	0.456	0.233	0.398	0.383	0.565	0.309	0.471	0.462
+ TTS-VAR ($N{=}1$)	0.773	0.690	0.543	0.269	0.424	0.391	0.596	0.311	0.515	0.500
+ TTS-VAR ($N{=}2$)	0.779	0.679	0.547	0.265	0.417	0.391	0.615	0.311	0.513	0.501
+ VISTA	**0.832**	**0.750**	**0.574**	**0.442**	**0.452**	**0.398**	0.572	0.310	**0.575**	**0.541**
Infinity-8B	0.827	0.753	0.604	0.365	0.414	0.397	0.612	0.316	0.560	0.536
+ VISTA	0.858	0.798	0.658	0.415	0.424	0.403	0.612	0.316	0.593	0.560

4 Experiments

4.1 Setup

Evaluation spans two Infinity model sizes and two complementary benchmarks. We evaluate VISTA on Infinity-2B and Infinity-8B Han et al. (2025a), using each model's default 13-scale schedule, and report results on T2I-CompBench Huang et al. (2023) and GenEval Ghosh et al. (2023). VISTA and its corresponding Infinity baselines are averaged over four seeds (Appendix F); BoN and TTS-VAR are single runs under our setup, and prior-system rows are reported from Shahabadi et al. (2025). Every Infinity row, baseline and VISTA alike, uses ScaleKV Li et al. (2026) cache compression. We refer to these configurations as **Infinity-2B** and **Infinity-8B** throughout and identify uncompressed results explicitly.

The setup isolates VISTA's contribution. Because VISTA operates on top of ScaleKV, the compressed backbone is its direct reference. Appendix F also reports the uncompressed Infinity-2B baseline: no category changes by more than 0.008, and the targeted average is essentially unchanged (0.481 vs. 0.480), so compression does not explain VISTA's gains. Unless stated otherwise, VISTA activates $\mathcal{L}_{\text{bind}} + \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{depth}}$ and steers the $4{\times}4$, $6{\times}6$, and $8{\times}8$ scales with $N_{\max} = 5$. The base step size lr_{base} and gradient threshold τ_k are set per backbone; Appendix J reports their values and selection details. No numeracy or non-spatial objective is active. We therefore report those categories unchanged from the corresponding baseline and exclude them from the targeted average.

4.2 Main Results

Table 1 reports T2I-CompBench results for both backbones and situates them against four diffusion and diffusion-transformer systems from prior benchmarking work Shahabadi et al. (2025).

VISTA improves every targeted category on both model sizes. On Infinity-2B, the largest gain is on 2D Spatial ($+88.9\%$), followed by improvements of 11.1–20.8% on Color, Texture, and Shape, 11.6% on 3D Spatial, and 3.6% on Complex. Complex improves least because its prompts combine several constraints and no single objective targets their interaction. Across the six targeted categories, VISTA raises the mean by 19.8% on 2B and 5.9% on 8B. The smaller relative gain on 8B reflects its stronger starting point and reduced headroom. The targeted-average standard deviation is 0.002 on both model sizes, compared with gains of 0.095 and 0.033, respectively (Appendix F).

Per-category results isolate the three instantiated objectives. VISTA activates a term only when the prompt contains its corresponding constraint. Color, Texture, and Shape therefore isolate $\mathcal{L}_{\text{bind}}$,

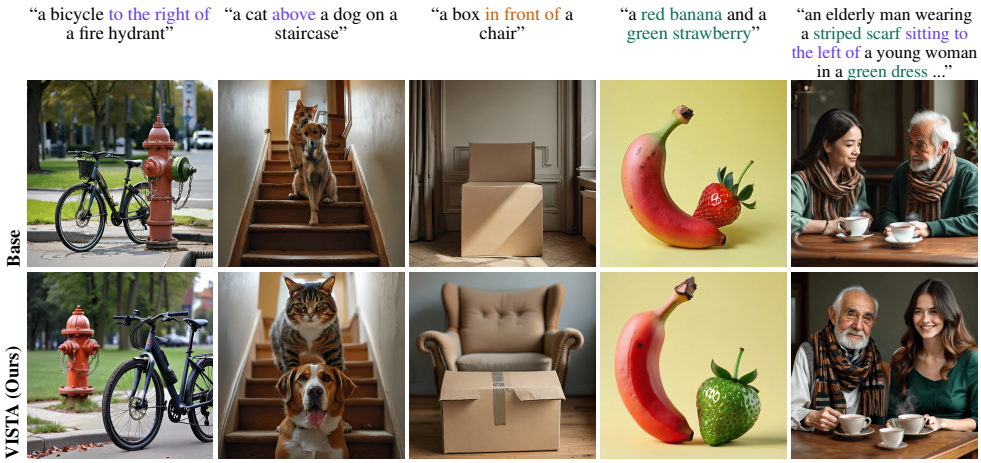

Figure 2: Matched Base and VISTA generations across planar relations, depth ordering, attribute binding, and their combination. Purple marks planar relations; orange marks depth relations; teal marks attribute–noun bindings.

2D Spatial isolates $\mathcal{L}_{\text{2D}}$, and 3D Spatial isolates $\mathcal{L}_{\text{depth}}$. These columns provide objective-specific evidence while preserving the complete-system evaluation on Complex prompts.

Test-time alignment lets 2B compete with a $4\times$ larger backbone. VISTA on Infinity-2B surpasses unmodified Infinity-8B on four of the six targeted categories — Color, 2D Spatial, 3D Spatial, and Complex — and on their average (0.575 vs. 0.560). It trails only on Texture and Shape. This suggests that some of the observed compositional gap between the two model sizes can be recovered by steering the smaller model.

4.3 Qualitative Results

VISTA improves isolated and joint compositional constraints. Figure 2 compares Base and VISTA under matched prompts and random seeds. The first two examples isolate planar relations, the third isolates depth ordering, the fourth isolates attribute binding, and the fifth combines planar and binding constraints. In the box–chair example, Base omits the chair, whereas VISTA renders both objects with the box in front of the chair. In the combined example, Base places the man on the wrong side of the woman and leaks the striped and green attributes across subjects; VISTA corrects both constraints. The cat–dog example also shows less subject merging under VISTA, a beneficial side effect of better-separated attention maps.

Coarse corrections persist through unsteered scales. Figure 1 shows the mechanism behind these changes. The *flower* and *crow* maps are nearly coincident under Base, separate at the steered scales, and remain distinct through the unsteered finer scales. Correcting the coarse layout is sufficient because later scales inherit and refine it rather than reconstructing it. Appendix G provides additional depth-ordering examples.

4.4 GenEval

We evaluate the same configurations on GenEval to test whether the gains transfer beyond T2I-CompBench.

The gains transfer across benchmarks and model scales. Position improves by 198% on 2B (0.250 $\rightarrow$ 0.745) and 42.7% on 8B, raising the overall score by 13.8% and 5.9%, respectively. VISTA on 2B surpasses unmodified 8B on Position (0.745 vs. 0.578), while attribute binding improves on both backbones. Color decreases slightly on 2B (0.827 $\rightarrow$ 0.809), an effect absent on 8B and T2I-CompBench; GenEval's Color category measures single-object attribution rather than the multi-object binding targeted by $\mathcal{L}_{\text{bind}}$. SSG reports a $+0.007$ GenEval Overall change on Infinity-

Table 2: GenEval over four seeds. All rows use ScaleKV, and VISTA uses the configuration of Table 1. Shaded categories are untargeted and unchanged by construction. **Bold**: best within each backbone group; underlined: best overall.

Model	Color	Attr.	Pos.	Single	Two	Count	Overall
Infinity-2B	0.827	0.560	0.250	0.994	0.808	0.591	0.672
+ VISTA	0.809	**0.645**	**0.745**	0.994	0.808	0.591	**0.765**
Infinity-8B	0.886	0.765	0.578	1.000	0.937	0.778	0.824
+ VISTA	0.907	0.788	0.825	1.000	0.937	0.778	0.873

Table 3: Quality–cost trade-off as the number of steered scales increases coarse-to-fine from 4×4. The fidelity columns report ImageReward (prompt-conditioned) and LAION aesthetic v2 (unconditional) on 600 paired prompts; Appendix I reports additional scorers. The shaded row is our default.

Steered scales	Targ. 2B ↑	Targ. 8B ↑	ImageReward 2B ↑	Aesthetic 2B ↑	Time 2B ↓
None	0.480	0.560	1.025	5.734	3.34 s
1	0.546	0.579	1.234	5.683	3.99 s
2	0.567	0.587	1.234	5.615	4.74 s
3	0.575	0.593	1.230	5.609	5.46 s
4	0.579	0.586	1.204	5.608	6.28 s

2B, while VPG reports $+0.01$ Overall and $+0.02$ Position Shin et al. (2026); Liao et al. (2026). Because their baselines and prompt protocols differ (VPG uses rewritten prompts, whereas ours remain unchanged), Appendix L compares their within-paper changes separately rather than inserting unmatched rows here.

Steering helps when sampling rarely contains a correct candidate. Table 1 compares VISTA with BoN and TTS-VAR Chen et al. (2025). Following TTS-VAR, BoN generates multiple complete trajectories and selects the final image with the highest ImageReward score Xu et al. (2023). This is effective when a satisfactory generation already appears with appreciable probability, but increasing the sample count offers limited help when the backbone repeatedly makes the same compositional error. The distinction is clearest for spatial relations: neither BoN nor increasing TTS-VAR from $N{=}1$ to $N{=}2$ provides consistent improvement, whereas VISTA steers the trajectory toward the requested arrangement. Sampling performs better on Numeracy, where successes vary across trajectories and VISTA activates no corresponding objective.

Candidate search introduces a different resource trade-off. TTS-VAR maintains multiple trajectories under a descending batch schedule, using DINOv2 clustering at coarse scales and ImageReward-based resampling later. Its computation and memory grow with search width and require intermediate decoding through additional pretrained models. VISTA instead optimizes one trajectory from internal cross-attention, requiring neither a candidate batch nor an external verifier. Its overhead depends on the number of steered scales and optimization steps rather than the number of sampled trajectories.

4.5 Quality, Fidelity, and Cost

Table 3 varies the number of steered scales along three axes: compositional accuracy, image fidelity, and generation time.

Three scales provide the strongest quality–cost balance. The targeted average saturates on both model sizes: a fourth scale raises runtime by 15.0%, improves 2B by only 0.004, and reduces 8B by 0.007. At the default setting, ImageReward increases by about 20% and CLIPScore by 1.1%, whereas the two LAION aesthetic predictors decrease by 1.3–2.2% (Appendix I). Three scales is therefore our default; reducing the step ceiling to $N_{\max}{=}1$ lowers three-scale runtime to 3.71 s. Appendix J provides the complete runtime and memory breakdown.

Table 4: Human evaluation on 200 paired generations. Counts are plurality-aggregated outcomes, one per image pair; $\text{Win}_{\text{dec.}}$ is VISTA's share after excluding ties. The p values use two-sided exact binomial tests over decisive outcomes. Appendix H reports the protocol, category breakdown, and annotator agreement.

Criterion	Base	**VISTA**	Tie	$\text{Win}_{\text{dec.}}$	p
Compositionality	13	**65**	122	**83.3%**	1.8×10^{-9}
Perceived quality	45	52	103	53.6%	0.543

Human judgments support compositional improvements across constraint types. Annotators evaluated 200 paired generations presented in randomized order without method identities. As Table 4 shows, VISTA wins 83.3% of decisive pair-level compositionality comparisons ($p < 10^{-8}$). The same pattern appears descriptively within every targeted category: VISTA receives 75.5% of decisive annotation-level preferences for attribute binding, 92.8% for 2D spatial relations, and 83.8% for 3D spatial relations. The compositional judgments support the benchmark gains. Separating attention maps may also reduce object and attribute merging, as the qualitative examples illustrate. Annotators show no significant overall preference between Base and VISTA in perceived quality ($p = 0.543$). Appendix H reports the complete protocol, counts, category breakdown, and annotator agreement.

5 Limitations and Future Work

Efficiency and backbone generality. VISTA trades additional inference time for compositional control, although Sec. 4.5 shows that this trade-off is adjustable. We evaluate two model sizes, but both belong to the Infinity family. The framework is designed around three general properties of next-scale VAR generation: a continuous input at each scale, persistent state across scales, and discrete sampling; but validation on other VAR backbones remains important. Future work can also reduce overhead through more selective scale and step allocation.

Broader compositional coverage remains open. We address attribute binding, planar relations, and depth ordering. Numeracy fits the framework but is disabled here, while non-spatial relations require objectives we do not yet instantiate. Complex prompts gain least because they combine several constraints, whereas our objectives treat each constraint separately. Future work can extend the registry with numeracy and non-spatial objectives and develop joint constraints that capture interactions among compositional phenomena. Finally, attribute–noun pairs and relations are obtained from a dependency parse that handles the benchmark prompts and longer examples we examined, but more robust parsing is needed for linguistically diverse prompts.

6 Conclusion

VISTA brings gradient-based test-time alignment to next-scale visual autoregressive generation. It separates general-purpose optimization mechanisms from a pluggable registry of compositional objectives, allowing constraints to change without modifying the underlying framework. Instantiated for attribute binding, planar relations, and depth ordering, VISTA improves every targeted T2I-CompBench category on both model sizes and produces large gains in GenEval positioning and attribute binding. Human evaluation confirms stronger compositionality without a significant change in perceived quality. On T2I-CompBench, VISTA also enables a 2B-parameter model to surpass a model four times its size on most targeted categories and on their average.

Cross-attention provides a transferable interface for test-time alignment. VISTA shows that the central premise of diffusion-based alignment — using a frozen model's cross-attention as a differentiable, training-free handle on compositional structure — extends to stateful, multi-scale, discrete-token generation. Whereas existing VAR test-time methods select among sampled trajectories and depend on a satisfying candidate appearing in the sample set, VISTA modifies the trajectory itself toward the requested constraints. By making attention-based objectives usable inside VAR sampling, VISTA opens next-scale autoregressive generators to a broad class of training-free compositional interventions.

REPRODUCIBILITY STATEMENT

We provide the implementation used for our experiments as supplementary material. Sec. 4.1 specifies the backbones, benchmarks, baselines, and default configuration. Secs. 3.2–3.4 describe the attention capture, objective interface, optimization, and scheduling mechanisms. Appendix H details the human-evaluation protocol, Appendix C gives the complete depth-objective construction and hyperparameters, and Appendices F and J report seed variance and system measurements. Appendices K and L document the comparison protocols, while Appendix B records the definitions and provenance of the instantiated objectives. Appendices I, D, and E provide the fidelity protocol, step-budget rule, and alternative objective instantiations. Together with the supplementary code, these details support reproduction of the reported results.

AI USE DISCLOSURE

In this work, we used generative AI tools to formulate mathematical claims and assist with method implementation. We did not use generative AI tools to develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results. Generating synthetic datasets, translation, dataset cleaning or reformatting, and qualitative or thematic data analysis were not applicable to this work.

Additionally, we used generative AI tools to draft parts of the paper and to aid and polish the writing. We reviewed all AI-assisted work. AI-assisted mathematical claims were manually verified, and AI-assisted implementation was reviewed and tested by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2283–2293. IEEE, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Zhekai Chen, Ruihang Chu, Yukang Chen, Shiwei Zhang, Yujie Wei, Yingya Zhang, and Xihui Liu. Tts-var: A test-time scaling framework for visual auto-regressive generation. *arXiv preprint arXiv:2507.18537*, 2025.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

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, 2025a.

Woojung Han, Yeonkyung Lee, Chanyoung Kim, Kwanghyun Park, and Seong Jae Hwang. Spatial transport optimization by repositioning attention map for training-free text-to-image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18401–18410. IEEE, 2025b.

Harold Haodong Chen, Xianfeng Wu, Wen-Jie Shu, Rongjin Guo, Disen Lan, Harry Yang, and Ying-Cong Chen. Go with your gut: Scaling confidence for autoregressive image generation. *arXiv e-prints*, pp. arXiv–2509, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kunjun Li, Zigeng Chen, Cheng-Yen Yang, and Jenq-Neng Hwang. Memory-efficient visual autoregressive modeling with scale-aware kv cache compression. *Advances in Neural Information Processing Systems*, 38:13736–13761, 2026.

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864*, 2023.

Xinyao Liao, Qiyuan He, Yicong Li, Jiayin Zhu, Xiaoye Qu, Wei Wei, and Angela Yao. Vpg: Visual prefix guidance for autoregressive image and video generation. *arXiv preprint arXiv:2605.30317*, 2026.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9005–9014. IEEE, 2024.

Joonhyung Park, Hyeongwon Jang, Joowon Kim, and Eunho Yang. Progress by pieces: Test-time scaling for autoregressive image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38091–38100, 2026.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023.

Parham Rezaei, Arash Marioriyad, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Why settle for mid: A probabilistic viewpoint to spatial relationship alignment in text-to-image models. *arXiv preprint arXiv:2506.23418*, 2025.

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.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Hossein Shahabadi, Niki Sepasian, Arash Marioriyad, Ali Sharifi-Zarchi, and Mahdieh Soleymani Baghshah. Infinity and beyond: Compositional alignment in var and diffusion t2i models. *arXiv preprint arXiv:2512.11542*, 2025.

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

Weidong Tang, Xinyan Wan, Siyu Li, and Xiumei Wang. Inference-time scaling for visual autoregressive modeling by searching representative samples. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 406–419. Springer, 2025.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Xiaohang Zhan and Dingming Liu. Larender: Training-free occlusion control in image generation via latent rendering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19679–19688. IEEE, 2025.

APPENDIX

A RELATED WORK

Sec. 1 positions VISTA between two families of test-time methods: gradient-based alignment for diffusion and inference-time search or guidance for autoregressive generation. We review both more fully here and identify the prior methods underlying our instantiated objectives.

Diffusion methods establish cross-attention as a test-time control interface. Attend-and-Excite Chefer et al. (2023) optimizes the diffusion latent during sampling to strengthen attention to neglected subject tokens. SynGen Rassin et al. (2023) uses a syntactic parse to align attributes with their nouns while separating them from unrelated concepts. CONFORM Meral et al. (2024) casts this alignment as a multi-positive contrastive objective over attention-map embeddings; we use this formulation for attribute binding. Divide&Bind Li et al. (2023) adds an attendance objective for multi-instance prompts, while A-STAR Agarwal et al. (2023) discourages overlap between concept maps, providing the segregation principle behind our separation term. For planar relations, PSG Rezaei et al. (2025) introduces the Probability-of-Superiority objective used by VISTA, while Han et al. (2025b) reposition attention mass through optimal transport.

Diffusion steering does not directly fit next-scale VAR generation. These methods share VISTA's premise that a frozen model's cross-attention provides a differentiable, training-free handle on compositional structure. However, they optimize a continuous, fixed-resolution latent that is repeatedly updated across denoising steps. Next-scale VAR models instead generate discrete tokens over increasingly fine grids, with each scale's representation derived and consumed once while persistent state is carried forward. Secs. 3.2–3.4 provide the optimization mechanisms needed to use the same attention-based objectives in this setting.

Autoregressive test-time methods search or guide. TTS-VAR Chen et al. (2025) casts next-scale generation as a path-search problem, combining clustering-based diversity preservation at coarse scales with reward-based selection at finer scales. GridAR Park et al. (2026) progressively searches and prunes partial raster-scan generations, while ScalingAR Haodong Chen et al. (2025) allocates test-time computation in next-token generation using token-entropy signals. A complementary line modifies next-scale logits without candidate search: SSG Shin et al. (2026) amplifies scale-specific semantic residuals obtained from a frequency-derived coarse prior, whereas VPG Liao et al. (2026) contrasts predictions under genuine and corrupted visual prefixes to reduce prefix drift. These methods target candidate quality, scale consistency, or prefix support, but do not explicitly optimize prompt-derived binding and spatial constraints. VISTA instead differentiates compositional objectives through cross-attention to modify a single coarse-scale representation. Table 1 compares VISTA directly with TTS-VAR and BoN, and Appendix L gives a protocol-aware comparison with SSG and VPG.

Depth ordering can be approached through different spatial cues. For 3D relations, PSG Rezaei et al. (2025) searches over completed images using segmentation and monocular-depth estimators. LaRender Zhan & Liu (2025) instead applies volume-rendering principles in diffusion latent space using object regions and an occlusion order. Our proposed depth objective provides a complementary attention-only formulation based on the prompt relation and the backbone's cross-attention, without requiring a depth estimator or supplied layout. It is one objective instantiated within VISTA rather than a requirement of the framework.

VISTA builds on next-scale visual autoregressive generation. VAR Tian et al. (2024) introduces coarse-to-fine next-scale prediction, and Infinity Han et al. (2025a) extends it to text-conditioned generation with a bitwise tokenizer and self-correction. Prior benchmarking work Shahabadi et al. (2025) documents Infinity's compositional failures relative to diffusion and diffusion-transformer systems and provides the prior-system results in Table 1. ScaleKV Li et al. (2026) reduces Infinity's cache cost by allocating different KV budgets across layers and scales. We apply it to all Infinity configurations, with Appendix F showing that its effect on the reported compositional scores is negligible.

The benchmarks expose complementary aspects of compositionality. T2I-CompBench Huang et al. (2023) evaluates attribute binding, spatial and non-spatial relations, counting, and complex compositions through category-specific prompts and metrics. Its separation of constraint types lets individual categories isolate the corresponding VISTA objectives. GenEval Ghosh et al. (2023) provides a complementary object-centric evaluation of presence, counting, color attribution, attribute binding, and relative position. Together, they expose recurring failures across model families, including attribute leakage, missing objects, incorrect arrangements, and miscounting.

B Objective Definitions and Provenance

This appendix specifies the binding and planar objectives used by VISTA and distinguishes components adopted from prior work from those introduced here. Both objectives operate on the token-group attention map $\mathbf{m}_T$ defined in Sec. 3.3.

Attribute binding. Following attention-based binding methods Rassin et al. (2023); Meral et al. (2024), let $p = (a, n) \in \mathcal{P}$ denote an attribute–noun pair, where a and n are the token groups corresponding to the modifier and its noun. For example, "a red banana" gives $p = (\mathsf{red}, \mathsf{banana})$. For each $T \in \{a, n\}$, we reshape $\mathbf{m}_T$ to the spatial grid, smooth it with a fixed 3×3 Gaussian kernel G with $\sigma = 0.5$, and ℓ_2-normalize the flattened result:

$$\mathbf{u}_T = \frac{\mathrm{vec}(G * \mathrm{reshape}(\mathbf{m}_T))}{\|\mathrm{vec}(G * \mathrm{reshape}(\mathbf{m}_T))\|_2}. \tag{9}$$

The resulting $M = 2|\mathcal{P}|$ embeddings form a contrastive objective adapted from CONFORM Meral et al. (2024). For an embedding i, $P(i)$ contains the embedding associated with the other member of the same attribute–noun pair, while embeddings from other pairs act as negatives. With $\mathrm{sim}_{ij} = \mathbf{u}_i^\top \mathbf{u}_j/\eta$ and temperature $\eta = 0.07$,

$$\mathcal{L}_{\mathrm{bind}}(\mathcal{P}) = -\frac{1}{M}\sum_{i=1}^{M}\frac{1}{|P(i)|}\sum_{j\in P(i)}\log\frac{\exp(\mathrm{sim}_{ij})}{\sum_{j'\neq i}\exp(\mathrm{sim}_{ij'})}. \tag{10}$$

The loss is evaluated jointly over all extracted pairs, encouraging each attribute map to align with its corresponding noun while remaining distinguishable from competing pairs.

Planar relations. For a relation $r = (X, \rho, Y) \in \mathcal{R}_{2\mathrm{D}}$ along image-plane axis a, we use the probability-of-superiority formulation from PSG Rezaei et al. (2025). Each attention map is first converted into a spatial distribution,

$$p_Z = \mathrm{softmax}(\gamma \mathbf{m}_Z), \qquad Z \in \{X, Y\}, \tag{11}$$

where γ sharpens the diffuse attention maps encountered at coarse VAR scales. The probability that X lies above Y in the ordering induced by axis a is

$$\mathrm{PoS}_a(X, Y) = \sum_{i,j} p_X(i)\, p_Y(j)\, \mathbb{1}[\mathrm{coord}_a(i) \geq \mathrm{coord}_a(j)]\,. \tag{12}$$

For relations aligned with the increasing coordinate direction, we set $q_r = \mathrm{PoS}_a(X, Y)$; for the opposite direction, such as *left of* or *above*, we use $q_r = 1 - \mathrm{PoS}_a(X, Y)$. The planar objective is

$$\mathcal{L}_{2\mathrm{D}} = -\frac{1}{|\mathcal{R}_{2\mathrm{D}}|}\sum_{r\in\mathcal{R}_{2\mathrm{D}}} q_r^2, \tag{13}$$

where the probability-of-superiority form and its squared objective follow PSG.

Provenance. Table 5 summarizes the origin of each objective component. We distinguish adopting a motivating principle from adopting a functional form.

Table 5: Provenance of the objective terms used in VISTA.

Term	Eq.	Prior work	VISTA-specific component
$\mathcal{L}_{\text{bind}}$	10	Contrastive binding formulation adapted from CONFORM Meral et al. (2024)	Application to grouped attention maps at coarse VAR scales
$\mathcal{L}_{\text{2D}}$	12	Probability-of-superiority formulation and squared objective from PSG Rezaei et al. (2025)	Temperature sharpening and directional implementation on discrete VAR grids
$\mathcal{L}_{\text{sep}}$	16	Segregation principle motivated by A-STAR Agarwal et al. (2023)	Extent-based functional form
$\mathcal{L}_{\text{bc}}$	22	—	Boundary-containment objective
$\mathcal{L}_{\text{entity}}$	—	—	Not instantiated
$\mathcal{L}_{\text{ns}}$	—	—	Not instantiated

We claim no new objective formulation for attribute binding or planar relations. The objective-level contribution is the depth term $\mathcal{L}_{\text{depth}} = \mathcal{L}_{\text{sep}} + \mathcal{L}_{\text{bc}}$: $\mathcal{L}_{\text{sep}}$ adopts the segregation motivation of A-STAR but uses a different extent-based function, while $\mathcal{L}_{\text{bc}}$ is introduced in this work. Unlike approaches that evaluate completed candidates using segmentation and monocular-depth models, this objective extracts a differentiable occlusion cue directly from the backbone's cross-attention maps and requires no auxiliary model.

C Depth Objective: Construction Details

This section specifies the depth objective summarized in Sec. 3.3. For a relation "B behind A," A denotes the foreground object and B the occluded object. We convert their attention maps into spatial distributions,

$$p_X = \text{softmax}(\gamma \mathbf{m}_X), \qquad X \in \{A, B\}, \tag{14}$$

defined over normalized grid coordinates $\mathbf{c}_i \in [0,1]^2$. Because cross-attention provides no viewing-axis coordinate, the objective uses overlap and boundary-containment cues associated with occlusion rather than estimating metric depth.

Separation. For each axis $a \in \{x, y\}$, we compute the attention-weighted centroid and spread

$$\mu_a^X = \sum_i p_X(i) c_{i,a}, \qquad \varrho_a^X = \sqrt{\sum_i p_X(i) \big(c_{i,a} - \mu_a^X\big)^2}. \tag{15}$$

The separation term penalizes the intersection of the resulting axis-aligned extents:

$$\mathcal{L}_{\text{sep}} = \prod_{a \in \{x,y\}} \left[\min_{X \in \{A,B\}} \big(\mu_a^X + \varrho_a^X\big) - \max_{X \in \{A,B\}} \big(\mu_a^X - \varrho_a^X\big) \right]_+, \tag{16}$$

where $[u]_+ = \max(u, 0)$. This discourages the two entities from collapsing into the same attention support.

Facing boundary. Separation alone does not assign foreground–background order. We approximate the perimeter of B using the spatial gradient magnitude of its reshaped attention distribution:

$$e_B(i) = \sqrt{\big(D_x p_B(i)\big)^2 + \big(D_y p_B(i)\big)^2}. \tag{17}$$

The derivatives use central finite differences in the grid interior and one-sided differences at its boundaries. We then emphasize the portion of this perimeter facing A:

$$s_i = e_B(i) \exp\left(- \frac{\big\| \mathbf{c}_i - \text{sg}(\boldsymbol{\mu}^A) \big\|_2}{\kappa} \right), \qquad \boldsymbol{\mu}^A = (\mu_x^A, \mu_y^A), \tag{18}$$

where sg denotes stop-gradient. This prevents the facing-selection term from being reduced simply by moving the centroid of A.

Virtual boundary points. Let $\mathcal{I}_K$ contain the K locations with the largest s_i, or all locations when the grid contains fewer than K, and define

$$w_k = \frac{s_k}{\sum_{\ell \in \mathcal{I}_K} s_\ell}. \tag{19}$$

For $\boldsymbol{\alpha}^{(v)} \sim \mathrm{Dir}(\mathbf{1}_{|\mathcal{I}_K|})$, we construct

$$\widetilde{\mathbf{x}}_v = \sum_{k \in \mathcal{I}_K} \frac{\alpha_k^{(v)} w_k}{\sum_{\ell \in \mathcal{I}_K} \alpha_\ell^{(v)} w_\ell} \mathbf{c}_k, \qquad v = 1, \ldots, V. \tag{20}$$

These virtual points lie in the convex hull of the selected boundary locations and concentrate on the portion of B facing A.

Boundary containment. We evaluate the attention density of A at each virtual point using a Gaussian kernel:

$$d_A(\widetilde{\mathbf{x}}_v) = \sum_i p_A(i) \exp\left(-\frac{\|\mathbf{c}_i - \widetilde{\mathbf{x}}_v\|_2^2}{2h^2} \right). \tag{21}$$

The containment loss requires these points to fall within sufficiently strong support from the foreground object:

$$\mathcal{L}_{\mathrm{bc}} = \frac{1}{V} \sum_{v=1}^{V} \left[\delta - d_A(\widetilde{\mathbf{x}}_v) \right]_+ . \tag{22}$$

We set δ relative to the peak attention of A, adapting the containment threshold to the concentration of its map.

Combined objective. The complete objective is

$$\mathcal{L}_{\mathrm{depth}} = \lambda_{\mathrm{sep}} \mathcal{L}_{\mathrm{sep}} + \lambda_{\mathrm{bc}} \mathcal{L}_{\mathrm{bc}}. \tag{23}$$

The separation term preserves distinct entity supports, while boundary containment assigns their foreground–background order through an occlusion cue. The objective does not recover metric depth or a general 3D representation.

All experiments use $\gamma = 8$, $K = 64$, $V = 256$, $\kappa = 0.3$, $h = 0.15$, $\delta = 0.3 \max_i p_A(i)$, and $\lambda_{\mathrm{sep}} = \lambda_{\mathrm{bc}} = 1$. Alternative depth objectives under the same VISTA machinery are evaluated in Appendix E.

D Adaptive Step-Budget Scheduler

For each objective active at the initial step, we define

$$\varsigma_k = \mathrm{clip}\left(\frac{\log\left(\overline{g}^{(k,0)}/\tau_k\right)}{\log(N_{\max} + 1)}, 0, 1 \right), \qquad N_{\mathrm{eff}} = \max\left(1, \left\lceil N_{\max} \max_{k \in \mathcal{K}_0} \varsigma_k \right\rceil \right). \tag{24}$$

The clipped log ratio maps the initial constraint violation to a bounded step budget while reducing sensitivity to large differences in gradient magnitude. The maximum allocates the budget according to the most violated active objective, and the outer maximum ensures that every selected scale receives at least one update. N_{eff} is only an upper bound: optimization terminates earlier when all objectives fall below their activity thresholds. We use this scheduler as a fixed compute-control heuristic throughout the experiments.

E Alternative Objective Instantiations

VISTA separates its optimization machinery from the objectives applied to the captured cross-attention maps. To test this interface, we hold the capture, normalization, combination, and budget mechanisms fixed and substitute alternative binding, planar, and depth objectives.

Each instantiation is evaluated with one seed on a fixed subset of 100 prompts from its corresponding T2I-CompBench category. Color, Texture, and Shape isolate the binding objective; 2D Spatial isolates the planar objective; and 3D Spatial isolates the depth objective. Complex prompts are omitted because they activate multiple terms simultaneously. The unsteered baselines are recomputed on the same subsets, so these diagnostic results should not be compared row by row with the full-benchmark results in Table 1. The experiment tests compatibility with the shared interface rather than providing a statistically powered ranking of objectives.

Table 6: Binding objectives on the three categories where only $\mathcal{L}_{\text{bind}}$ is active. All rows share the same optimization machinery and the same planar and depth terms, which are inactive here. Δ is on the mean.

$\mathcal{L}_{\text{bind}}$	Color	Texture	Shape	Mean	Δ
CONFORM (default)	0.830	0.764	0.570	0.721	$+17.1\%$
SynGen divergence	0.823	0.745	0.565	0.711	$+15.5\%$
KL	0.830	0.740	0.556	0.709	$+15.1\%$
Focal	0.769	0.735	0.579	0.694	$+12.8\%$
Divide-and-Bind	0.778	0.633	0.459	0.623	$+1.2\%$
JSD	0.730	0.642	0.475	0.616	$+0.0\%$
No steering	0.730	0.642	0.475	0.616	—

Table 7: Planar objectives on 2D Spatial, where only $\mathcal{L}_{\text{2D}}$ is active.

$\mathcal{L}_{\text{2D}}$	2D Spatial	Δ
Relative centroid	0.473	$+175.9\%$
PoS (default)	0.454	$+164.8\%$
Gaussian-at-target	0.417	$+143.3\%$
Half-field mass	0.363	$+111.8\%$
Separation	0.225	$+31.3\%$
Attention IoU	0.198	$+15.5\%$
No steering	0.171	—

Table 8: Depth objectives on 3D Spatial, where only $\mathcal{L}_{\text{depth}}$ is active.

$\mathcal{L}_{\text{depth}}$	3D Spatial	Δ
BC (ours)	0.495	$+12.8\%$
Peak dominance	0.487	$+11.0\%$
Overlap dominance	0.478	$+8.9\%$
Containment–adjacency	0.458	$+4.3\%$
Depth proxy	0.419	-4.5%
No steering	0.439	—

Across the 17 tested instantiations, 15 improve over their corresponding unsteered baseline, one matches it, and one underperforms. Thus, objectives with substantially different formulations can use the same capture and optimization machinery without framework-specific modifications. This supports VISTA's objective-agnostic interface; it does not imply that every possible attention objective will be equally effective.

A descriptive pattern is that objectives operating on spatial summaries transfer more reliably at coarse scales than objectives based primarily on elementwise map comparisons. Boundary containment performs best among the tested depth objectives. For planar relations, relative centroid slightly exceeds the literature-derived PoS objective on this diagnostic subset. We retain PoS as the default to keep the main configuration tied to the established objective rather than selecting a replacement retrospectively on this small study. Given the single seed and subset size, we draw no conclusions from small differences between the stronger variants.

F Per-Seed Results and Variance

Table 9 reports per-seed T2I-CompBench results for VISTA's default configuration on both backbones. VISTA and the corresponding Infinity baselines in Table 1 are averaged over four seeds. BoN and TTS-VAR use one seed, while results for other systems are taken from their respective papers.

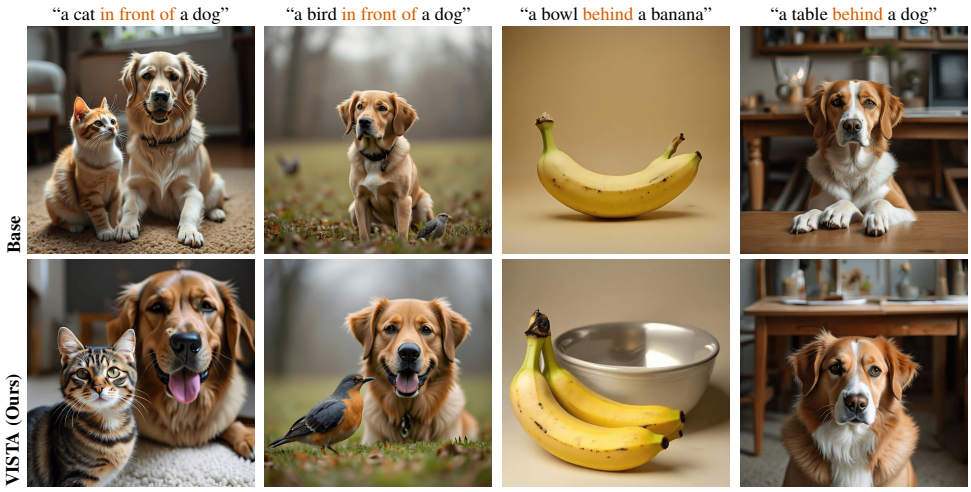

Figure 3: Selected qualitative comparisons for depth ordering. Each column uses the same prompt and random seed for Base and VISTA. The prompts contain only a front–back relation, isolating $\mathcal{L}_{\text{depth}}$ from the binding and planar objectives. Across these examples, VISTA produces clearer overlap and occlusion cues consistent with the requested ordering. Quantitative results over the complete 3D Spatial category appear in Table 1.

Table 9: Per-seed T2I-CompBench scores for VISTA's default configuration of three steered scales and $N_{\max}$=5. Avg. is the mean over the six targeted categories.

Model	Seed	Col.	Tex.	Shp.	2D	3D	Cpx.	Avg.
2B	0	0.826	0.750	0.566	0.458	0.455	0.394	0.575
	1	0.823	0.755	0.579	0.432	0.446	0.400	0.573
	2	0.853	0.744	0.576	0.439	0.453	0.399	0.577
	3	0.828	0.752	0.576	0.438	0.453	0.399	0.574
	Std.	0.014	0.005	0.006	0.011	0.004	0.003	0.002
8B	0	0.851	0.802	0.652	0.429	0.426	0.406	0.594
	1	0.863	0.790	0.666	0.417	0.424	0.402	0.594
	2	0.858	0.801	0.657	0.404	0.421	0.402	0.591
	3	0.860	0.800	0.656	0.411	0.425	0.402	0.592
	Std.	0.005	0.006	0.006	0.011	0.002	0.002	0.002

Variance across seeds. The targeted-average standard deviation is 0.002 on both backbones, and the largest category-level standard deviation is 0.014. The improvements in the principal targeted categories therefore remain stable across seeds, including the large gain on 2D Spatial. Complex has the smallest improvement and should be interpreted more cautiously; our conclusions do not rely on that category alone.

Effect of cache compression. All Infinity results in Table 1 use ScaleKV, since VISTA is applied on top of the compressed backbone. As a reference check, the uncompressed Infinity-2B baseline obtains a targeted average of 0.481, compared with 0.480 with ScaleKV, and no category changes by more than 0.008. The effect is comparable to the observed seed variation and negligible at the level of the reported aggregate results. We therefore use the compressed backbone consistently throughout.

G Additional Qualitative Results

Unlike planar relations, front–back order cannot be expressed by ordering two objects along an image-plane axis. The depth objective instead encourages separation and boundary-containment

cues that make the requested ordering visible through occlusion. Figure 3 illustrates this intended behavior; it does not assume that cross-attention encodes metric depth.

H Human Evaluation

Four annotators independently evaluated 200 paired Base–VISTA generations: 100 attribute-binding, 50 2D-spatial, and 50 3D-spatial prompts sampled from T2I-CompBench. Image order was randomized independently for each comparison, and annotators were unaware of the method identities. Each annotator made separate three-way judgments (VISTA, Base, or tie) for compositional faithfulness and visual quality. For each pair and criterion, we select the label with the highest vote count among VISTA, Base, and tie; equal highest counts produce a pair-level tie. Thus, a 2–1–1 split selects the two-vote label, while a 2–2 split produces a tie. We report annotation-level category results descriptively and use two-sided exact binomial tests over decisive pair-level outcomes.

Table 10: Human evaluation. Win denotes VISTA's share among decisive judgments. Category rows aggregate four annotations per pair; pair-level rows contain one plurality-aggregated outcome per image pair.

Criterion	n	VISTA	Base	Tie	Win
Compositional faithfulness					
Attribute binding	400	145	47	208	75.5%
3D spatial	200	57	11	132	83.8%
2D spatial	200	90	7	103	92.8%
All annotations	800	292	65	443	81.8%
Pair-level vote	200	65	13	122	83.3%
Visual quality					
All annotations	800	238	218	344	52.2%
Pair-level vote	200	52	45	103	53.6%

Agreement was moderate for compositional faithfulness (Fleiss' $\kappa = 0.589$) and fair for visual quality ($\kappa = 0.333$). For decisive pair-level outcomes, VISTA's compositional advantage is significant ($p = 1.8 \times 10^{-9}$), whereas visual quality shows no significant preference ($p = 0.543$).

I Prompt Fidelity and Visual Quality

We evaluate Infinity-2B and VISTA on 600 T2I-CompBench prompts, with 100 prompts from each targeted category. The random seed is held fixed, pairing every VISTA generation with a base generation from the same prompt and seed. We report three metric families that are not used by VISTA: ImageReward Xu et al. (2023), CLIPScore Hessel et al. (2021), and two released versions of the LAION aesthetic predictor Schuhmann et al. (2022). Statistical significance is assessed using a two-sided Wilcoxon signed-rank test over the 600 paired generations. This single-seed analysis complements the multi-seed compositional results in Table 1.

Table 11: Prompt-conditioned and aesthetic scores as the number of steered scales increases. All VISTA generations are paired with base generations using the same prompts and seeds. Differences and p values compare the default three-scale configuration with the base model.

Metric	Base	1 scale	2 scales	3 scales	4 scales	Δ at default	p
ImageReward	1.025	**1.234**	**1.234**	1.230	1.204	$+20.1\%$	3.6×10^{-7}
CLIPScore	0.321	0.325	**0.326**	0.325	0.324	$+1.1\%$	6.1×10^{-4}
Aesthetic (v1)	**6.056**	6.022	6.023	5.976	5.957	-1.3%	1.6×10^{-3}
Aesthetic (v2)	**5.734**	5.683	5.615	5.609	5.608	-2.2%	2.0×10^{-10}

Steering improves prompt-conditioned scores while preserving perceived quality. At the default three-scale configuration, ImageReward increases by 20.1% and CLIPScore by 1.1%, while

the two aesthetic predictors decrease modestly by 1.3% and 2.2%. These metric families measure different properties: ImageReward and CLIPScore condition on the prompt, whereas the aesthetic predictors score the image independently of it. The results therefore indicate substantially stronger prompt agreement with a small trade-off in predicted aesthetics. Importantly, the human evaluation in Table 4 finds no significant preference in perceived quality ($p = 0.54$), providing no evidence of a human-perceptible quality penalty at the default setting.

The improvement is not driven by a single category. At the default configuration, ImageReward increases in all six evaluated categories, with per-category gains ranging from $+0.034$ on Complex to $+0.351$ on Texture. This consistency indicates that the aggregate improvement is distributed across the benchmark rather than arising from one category.

J Inference Overhead: Full Breakdown

Table 12 reports per-image generation time on Infinity-2B over every combination of step budget, steered-scale count, and active objective subset. All 2B timing configurations use the same 20 prompts and an NVIDIA GeForce RTX 3090 (24 GB). The Infinity-8B experiments use an NVIDIA A100 (40 GB). The unmodified baseline averages 3.34 s per image. Peak memory is 10.96 GB for the baseline and all configurations with up to three steered scales, rising to 15.1–15.3 GB at four scales.

Table 12: Per-image generation time in seconds on Infinity-2B. B, S, and D denote the binding, spatial, and depth objectives. The unmodified baseline is 3.34 s.

Scales	B	S	D	B+S	B+D	S+D	B+S+D
$N_{\max} = 1$							
1	3.34	3.31	3.34	3.39	3.38	3.39	3.42
2	3.44	3.45	3.45	3.54	3.50	3.52	3.57
3	3.53	3.54	3.56	3.71	3.63	3.64	3.71
4	3.63	3.63	3.64	3.83	3.75	3.79	3.88
$N_{\max} = 5$							
1	3.71	3.71	3.72	3.94	3.85	3.89	3.99
2	4.17	4.14	4.16	4.63	4.46	4.51	4.74
3	4.62	4.48	4.48	5.28	5.06	4.92	5.46
4	5.11	4.61	4.62	6.03	5.74	5.08	6.28

Objectives compose sub-additively. At the default setting of three scales and $N_{\max}$=5, binding, spatial, and depth steering add 1.28 s, 1.14 s, and 1.14 s when run independently. Summing these standalone costs would add 3.56 s, whereas jointly optimizing all three adds 2.12 s, approximately 60% as much. This follows from Sec. 3.4: the active objectives share the same optimization passes and contribute to one joint update rather than running sequentially.

The fourth scale adds cost without a consistent gain. At the default step budget, moving from three to four scales increases runtime from 5.46 to 6.28 s, or 15.0%. The targeted average improves by only 0.004 on 2B and decreases by 0.007 on 8B. This supports using three scales as the default and treating the number of steered scales as an explicit quality–cost control.

Memory on the larger backbone. On Infinity-8B, peak memory is 22.7 GB with up to three steered scales and rises to 36.1 GB at four. Three scales therefore remain the practical default on both model sizes.

Backbone-specific hyperparameters. The base step size is $\mathrm{lr}_{\mathrm{base}} = 0.005$ on Infinity-2B and 0.045 on Infinity-8B. The gradient threshold is $\tau_k = 10^{-3}$ on 2B and 10^{-12} on 8B, reflecting the different gradient scales of the two backbones. At the 8B setting, the threshold is low enough that early stopping and adaptive budget reduction are effectively inactive, so each selected scale uses the full $N_{\max}$. These settings were selected through brief manual exploration on a fixed prompt subset without grid search.

K Detailed Comparison with TTS-VAR

TTS-VAR Chen et al. (2025) and VISTA improve frozen next-scale VAR generation through different mechanisms. TTS-VAR searches over multiple sampled trajectories and reallocates computation toward selected prefixes; VISTA instead optimizes the internal representation of one trajectory. This appendix provides the implementation and evaluation details underlying the comparison in Table 1.

TTS-VAR mechanism. For the 13-scale Infinity-2B model, TTS-VAR uses the descending batch schedule

$$\{8N, 8N, 6N, 6N, 6N, 4N, 2N, 2N, 2N, N, N, N, N\}.$$

At scales 2 and 5, it decodes the current states, extracts DINOv2 features, and uses K-Means++ to retain structurally diverse candidates as the batch contracts. At scales 6 and 9, it decodes the remaining states, scores them with ImageReward, and resamples their prefixes according to the resulting potential scores. TTS-VAR therefore changes which sampled prefixes receive further computation without directly optimizing their representations.

Evaluation protocol. For the controlled comparison in Table 1, we run the authors' public implementation at $N{=}1$ and $N{=}2$ using the same Infinity-2B backbone, T2I-CompBench prompts, and evaluation code as VISTA. These experiments use one seed. We do not directly substitute the authors' published T2I-CompBench numbers because their reported Infinity baseline and the evaluators used for Non-Spatial and Complex differ from ours. Their $N{=}8$ configuration begins with 64 coarse candidates and exceeded our available compute budget.

Search and steering address different failure regimes. Search is effective when successful generations occur within the sampled candidate set. This is reflected in TTS-VAR's published results: increasing the search budget substantially improves categories such as two-object generation and counting, where outcomes vary across trajectories. Its gains are smaller for positional relations, for which the reported GenEval score changes from 0.2020 for the backbone to 0.2172 at $N{=}8$; the authors likewise identify Position as a remaining limitation. Table 1 shows the same distinction under our evaluation protocol: sampling helps most on categories with variable outcomes, whereas VISTA produces its largest gains on spatial relations by directly modifying the trajectory toward the requested arrangement.

The methods intervene at different stages. TTS-VAR reports that reward scores from early intermediate images are weak predictors of final outcomes. It therefore preserves structural diversity through clustering at scales 2 and 5 and applies reward-guided selection only at scales 6 and 9, where score consistency is higher. VISTA requires neither a decoded intermediate image nor a pixel-space verifier: its cross-attention objectives provide a differentiable signal at the coarse scales where scene structure is being established. Search and steering are consequently compatible in principle, but we do not evaluate a combined method.

Resource profiles. TTS-VAR's computation and memory grow with its candidate schedule and additionally require intermediate decoding, DINOv2 feature extraction, and ImageReward inference. Its published analysis on an A800-80GB reports peak memory of 30.73 GB at $N{=}2$ and 46.40 GB at $N{=}4$, compared with 20.34 GB for the backbone; the reported timing excludes intermediate image decoding. The paper further reports that BoN with $N{=}8$ could not fit in one pass within 80 GB and was therefore executed in two passes.

VISTA maintains one trajectory and uses no auxiliary feature extractor or reward model, but incurs additional forward and backward passes through the backbone. Its cost is controlled by the number of steered scales and optimization steps, as detailed in Appendix J. Because the published TTS-VAR measurements use a different hardware and timing protocol, we compare the methods' resource scaling rather than claiming a direct wall-clock advantage.

L Comparison with Next-Scale Guidance

SSG Shin et al. (2026), VPG Liao et al. (2026), and VISTA are training-free inference-time methods evaluated on Infinity-2B, but they optimize different signals. SSG amplifies scale-specific residuals

obtained from a frequency-derived coarse prior, targeting refinement and fidelity. VPG contrasts predictions under generated and corrupted prefixes, targeting prefix drift. VISTA instead differentiates prompt-derived compositional objectives through cross-attention.

Neither SSG nor VPG reports T2I-CompBench under the protocol of Table 1. Their GenEval results also cannot be inserted as controlled rows in Table 2: SSG reports only the aggregate score, while VPG uses rewritten prompts and reports values to two decimal places. VISTA uses the original benchmark prompts and averages four seeds. We therefore compare only within-paper changes from each method's own baseline.

Table 13: Reported Infinity-2B GenEval changes. Each arrow compares a method with its own baseline. Parentheses show absolute changes.

Method	Position	Overall
SSG Shin et al. (2026)	—	$0.683 \rightarrow 0.690$ $(+0.007)$
VPG Liao et al. (2026)	$0.39 \rightarrow 0.41$ $(+0.02)$	$0.70 \rightarrow 0.71$ $(+0.01)$
VISTA	$0.250 \rightarrow 0.745$ $(+0.495)$	$0.672 \rightarrow 0.765$ $(+0.093)$

The reported changes reflect the methods' different targets. SSG and VPG modify the generic sampling distribution without encoding a requested relation, whereas VISTA directly optimizes the position constraint measured by GenEval. The larger Position change therefore supports explicit constraint steering, but is not a controlled claim that VISTA is uniformly better in general image quality.