

ISAC: TRAINING-FREE INSTANCE-TO-SEMANTIC ATTENTION CONTROL FOR IMPROVING MULTI-INSTANCE GENERATION

Anonymous authors

Paper under double-blind review

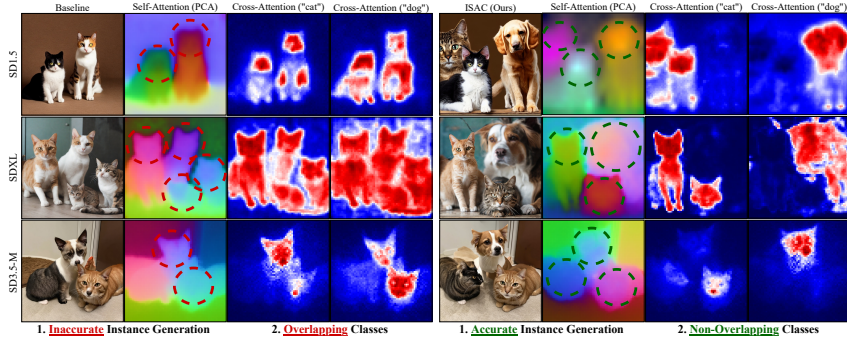


Figure 1: Comparison between existing text-to-image diffusion models (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024) (left) and the proposed ISAC framework (right) on the prompt "A photo of two cats and a dog". Text-to-image diffusion models struggle to clearly separate individual instances, leading to merged or overlapped objects (red dashed circles). In contrast, ISAC explicitly utilizes early-stage self-attention to accurately identify distinct, non-overlapping instances (green dashed circles) and assigns them precise semantic labels, resulting in clear class boundaries without additional training or supervision.

ABSTRACT

Text-to-image diffusion models excel at synthesizing single objects but frequently fail in multi-instance scenes, producing merged or missing objects. We show that this limitation arises because instance structures emerge before semantic features during denoising, making early semantic guidance unreliable. To address this, we propose Instance-to-Semantic Attention Control (ISAC), a training-free and hierarchical inference objective that first enforces non-overlapping instance formation with self-attention and then aligns semantics through cross-attention. ISAC introduces a maximum pixel-wise overlap (MPO) criterion to strictly decouple instances and can be applied either as latent optimization or latent selection. Experiments on T2I-CompBench, HRS-Bench, and a new similar-object benchmark show that ISAC substantially improves both multi-class and multi-instance fidelity, achieving up to 52% multi-class accuracy and 83% multi-instance accuracy without external supervision. Our findings highlight the importance of aligning control with diffusion dynamics for faithful and scalable multi-object generation. The code will be made available upon publication.

1 INTRODUCTION

Text-to-image (T2I) diffusion models (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024; Chen et al., 2023a; Labs, 2024) have demonstrated remarkable capabilities in generating high-quality images from textual descriptions. While these models excel at generating single objects, they often fail in multi-instance scenes, omitting instances or merging them (e.g., "A photo of two cats and a dog"; Figure 1).

Prior work attributes this issue to the failure of diffusion models to assign distinct spatial regions to individual instances. To address this, recent methods have attempted to enforce separation by utilizing semantic signals—either by manipulating cross-attention maps within the UNet (Chefer et al., 2023; Rassin et al., 2024; Guo et al., 2024; Hu et al., 2024; Meral et al., 2024; Qiu et al., 2025; Wang et al., 2024b; Jiang et al., 2024) or by adjusting per-instance text embedding (Feng et al., 2023; Hu et al., 2024; Chen et al., 2024a). However, these approaches are fundamentally limited because the semantic signals they rely on are unreliable and not yet well-formed early in denoising. Consequently, relying on these underdeveloped signals during this critical period, when the image’s core structure is established, often leads to the aforementioned failures.

In this paper, we discover that distinct semantics emerge *after* spatial instance structure forms, with diffusion dynamics analysis. Guided by this, we introduce Instance-to-Semantic Attention Control (ISAC), a *hierarchical, two-phase* objective: (1) form N non-overlapping instance structures from object counts; (2) bind semantics to those structures. Furthermore, to enable stricter separation than standard Intersection-over-Union (IoU), we propose a *maximum pixel-wise overlap (MPO)* criterion. We demonstrate the utility of the ISAC objective in two practical algorithms: latent optimization (using ISAC as a loss) and latent selection (using ISAC as a verifier).

We evaluate ISAC on the widely adopted T2I-CompBench (Huang et al., 2025), HRS-Bench (Bakr et al., 2023) benchmarks and an additional benchmark focused on similar-object scenarios. Across these settings, ISAC substantially improves multi-instance fidelity relative to recent approaches that do not account for diffusion dynamics. Because ISAC is model-agnostic, it also augments layout-guided pipelines. Qualitative results confirm these findings that ISAC remains highly beneficial even when instance locations are predefined.

To summarize, the key contributions of this work are:

- A novel analysis of the diffusion process for multi-instance generation that reveals a key temporal dynamic: distinct *instance structures* form early in the denoising process, well before cohesive *semantic features* emerge.
- Instance-to-Semantic Attention Control (ISAC), a novel *hierarchical, two-phase objective* designed around this dynamic, which prioritizes structural separation before enforcing semantic alignment.
- A new separation criterion, *maximum pixel-wise overlap (MPO)*, designed to enforce stricter instance boundaries than standard metrics like Intersection-over-Union (IoU).
- ISAC, as a *model-agnostic add-on*, consistently outperforms state-of-the-art methods on standard benchmarks and enhances existing layout-guided models.

2 RELATED WORK

A central challenge in text-to-image diffusion is the assignment of mutually exclusive spatial regions to multiple instances. Recent approaches seek to improve a model’s regional awareness and can be grouped into three categories.

Training-free Methods. Cross-attention has been established as the primary interface between textual semantics and spatial layout (Hertz et al., 2022). Building on this insight, several works manipulate cross-attention maps to encourage non-overlapping instance regions. Attend-and-Excite (Chefer et al., 2023) maximizes attention peaks for neglected objects, and InitNO (Guo et al., 2024) additionally incorporates self-attention maps to consider spatial features. On the other hand, SynGen (Rassin et al., 2024) and CONFORM (Meral et al., 2024) introduced a contrastive loss that aims to separate distinct instances while binding visual attributes to their corresponding instances, aided by a syntax parser (Honnibal, 2017). Self-Cross (Qiu et al., 2025) extends these works by additionally incorporating self-attention maps into contrastive guidance. A complementary line of works adjusts per-instance text embeddings or token ordering to influence attention behavior and thereby improve spatial separation (Feng et al., 2023; Hu et al., 2024; Chen et al., 2024a). Another approach introduces guidance from a pretrained vision model rather than relying solely on internal representations (Kang et al., 2025).

Despite these advances, training-free methods generally depend on semantic signals that are unreliable in the early denoising steps, precisely when global structure is being established. Since utilizing incorrect cross-attention maps during structure formation can lead to failure—and applying guidance

from pretrained models is ineffective after these structures are already formed—a new paradigm for instance control that considers diffusion dynamics is needed.

Fine-tuning Methods. Several approaches incorporate explicit segmentation signals through model adaptation. Methods such as TokenCompose (Wang et al., 2024b) and CoMat (Jiang et al., 2024) fine-tune the UNet so that its cross-attention maps align with segmentation masks obtained from a pretrained model (Ren et al., 2024). Alternatively, CountGen (Binyamin et al., 2024) fine-tunes the diffusion model to first generate segmentation masks and then uses them to guide the final image synthesis. However, a significant drawback of these techniques is their limited training vocabulary compared to the original models, which restricts their general applicability. The proposed ISAC objective is complementary to such methods and can be employed at inference time with or without fine-tuned components.

Layout-to-Image Methods. Recent systems adopt a two-stage pipeline in which a layout of bounding boxes is generated from a text prompt and then used to guide image synthesis (Lian et al., 2023; Zhang et al., 2024a). State-of-the-art controllers provide dense layout control (Li et al., 2023; Zhou et al., 2024a; Wang et al., 2024a; Cheng et al., 2024; Zhou et al., 2024b), yet they frequently struggle when objects are adjacent because explicit mechanisms for instance separation are absent. As a model-agnostic addition, ISAC directly enforces separation among neighboring instances and thereby improves the reliability of layout-conditioned generation.

3 PRELIMINARIES

We address the text-to-image (T2I) generation task using latent diffusion models. Given a text prompt C , which we assume provides a set of class tokens $\{\tau_i\}_{i=1}^k$ and their corresponding instance counts $\{n_i\}_{i=1}^k$, the model generates a corresponding image.

3.1 LATENT DIFFUSION MODELS

Latent Diffusion Models (LDMs) learn to reverse a forward noising process that gradually corrupts a clean image latent X_0 into random noise X_T . A neural network ϵ_θ is trained to predict the noise added at any timestep t , conditioned on the noisy latent X_t and a text embedding $\mathcal{T} \in \mathbb{R}^{L \times d}$. The model is optimized with the following objective: $\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \mathbb{E}_{(X_0, \mathcal{T}), \epsilon, t} [\|\epsilon_\theta(X_t, t, \mathcal{T}) - \epsilon\|_2^2]$.

Sample Generation. During inference, an image is synthesized with iterative denoising steps, $X_{t-1} \leftarrow \text{Denoise}(X_t, \mathcal{T}, \epsilon_\theta, t)$, starting from random noise $X_T \sim \mathcal{N}(0, I)$ to recover a clean latent \hat{X}_0 . A VAE decoder \mathcal{D} maps latent space to pixel space, obtaining a final image ($\hat{I} = \mathcal{D}(\hat{X}_0)$). Our framework is agnostic to generative dynamics (e.g., DDPM (Ho et al., 2020), Flow Matching (Esser et al., 2024)) and only requires access to the latent X_t and the model’s attention layers at each step.

Architecture and Attention. The denoiser ϵ_θ is typically a U-Net or Diffusion Transformer (DiT) containing multiple attention layers. The core of the denoiser uses two key attention mechanisms: *self-attention*, which captures spatial relationships within the image latent, and *cross-attention*, which aligns spatial features with the text embeddings.

Let $X_t \in \mathbb{R}^{HW \times d}$ be the reshaped latent and $\mathcal{T} \in \mathbb{R}^{L \times d}$ be the text embedding. For a given attention head, query (Q_t) and key (K_t) vectors are computed using learned projection matrices. For *self-attention*, both are derived from the latent: $Q_t^{\text{self}} = X_t W_Q^{\text{self}}$, $K_t^{\text{self}} = X_t W_K^{\text{self}}$, and then computed as $SA(X_t) = \text{softmax}(Q_t^{\text{self}} K_t^{\text{self}^\top} / \sqrt{d_h}) \in [0, 1]^{HW \times HW}$. For *cross-attention*, the query comes from the latent and the key from the text: $Q_t^{\text{cross}} = X_t W_Q^{\text{cross}}$, $K_t^{\text{cross}} = \mathcal{T} W_K^{\text{cross}}$, and then computed as $CA(X_t, \mathcal{T}) = \text{softmax}(Q_t^{\text{cross}} K_t^{\text{cross}^\top} / \sqrt{d_h}) \in [0, 1]^{HW \times L}$.

3.2 ATTENTION ACCUMULATION VIA HOOKS

During sampling we register forward hooks, $\mathcal{H}^{\text{self}}$ and $\mathcal{H}^{\text{cross}}$, on all attention layers to extract attention maps without altering the computation. Let there be M attention layers and h_l attention heads at the l -th layer. Since the denoiser contains attention maps with various spatial resolutions, we upsample the maps from each layer l and head h to the highest spatial resolution $H \times W$. These maps are then averaged to produce a single accumulated attention map for each type:

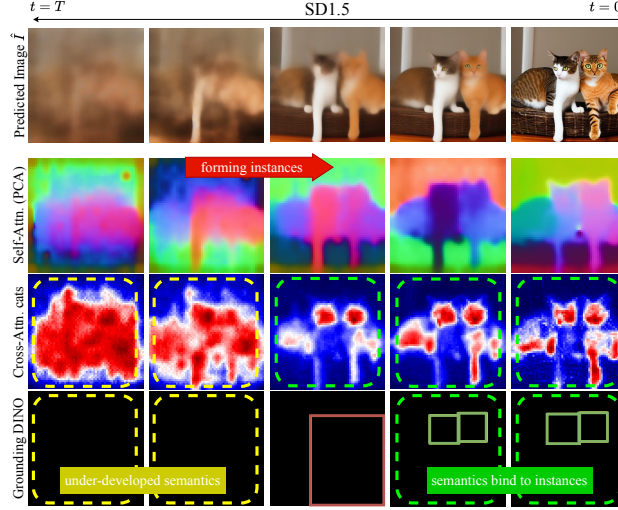


Figure 2: Temporal dynamics of diffusion models; Instance forms first, then semantic binds to it.

$$\mathcal{H}^{\text{self}}(X_t, \mathcal{T}, \epsilon_\theta, t) = \frac{1}{\sum_{l=1}^M h_l} \sum_{l=1}^M \sum_{h=1}^{h_l} \text{Upsample}(SA_l^h, \delta_l) \in \mathbb{R}^{HW \times HW} \quad (1)$$

$$\mathcal{H}^{\text{cross}}(X_t, \mathcal{T}, \epsilon_\theta, t) = \frac{1}{\sum_{l=1}^M h_l} \sum_{l=1}^M \sum_{h=1}^{h_l} \text{Upsample}(CA_l^h, \delta_l) \in \mathbb{R}^{HW \times L} \quad (2)$$

where $\delta_l = H/H_l = W/W_l$ is the upsampling factor for the l -th layer.

4 DISCOVERING DYNAMICS OF DIFFUSION MODELS IN MULTI-INSTANCE GENERATION

Figure 2 describes temporal dynamics of diffusion models. Before instance structures are formed, semantics of each instance is under-developed. Therefore, its internal representations, cross-attention maps, easily flood into the whole image. Also in this step, detection model (Liu et al., 2024) cannot find any instances because no recognizable semantic signals arise. When instance structures are formed and stabilized, their corresponding semantic signals bind to the instance and can also be recognizable by detection model. In Section E we show this is a model-agnostic behavior in the diffusion models family.

5 ISAC: DYNAMICS ALIGNED INSTANCE CONTROL OBJECTIVE

We propose a two-phase objective that first forms instance structures then binds semantics; it serves as a loss or a verifier. Figure 3 shows the overview of this objective.

5.1 PHASE 1: FORMING INSTANCE STRUCTURES WITH ONLY OBJECT COUNTS

When instance structures begin to form, semantic signals such as cross-attention maps from corresponding instances are still underdeveloped. Therefore, in Phase 1, instance separation must be performed using only structural signals. Suppose we target k classes, represented by class tokens τ_1, \dots, τ_k and a total of $N = n_1 + \dots + n_k$ instances. With only structural signals early on, we can conclude that “A total of N instances should occupy mutually exclusive, N distinct regions.”

To formulate this objective, we first create a *global foreground mask* to isolate relevant image regions. We then apply K-means clustering to the self-attention features within this foreground mask.

Details on obtaining a global foreground mask M_{fg} . Let $CA_t \in [0, 1]^{HW \times L}$, $SA_t \in [0, 1]^{HW \times HW}$ denote the accumulated cross- and self-attention map at timestep t , respectively, where L is the length of the prompt token sequence. Following (Binyamin et al., 2024), we obtain a

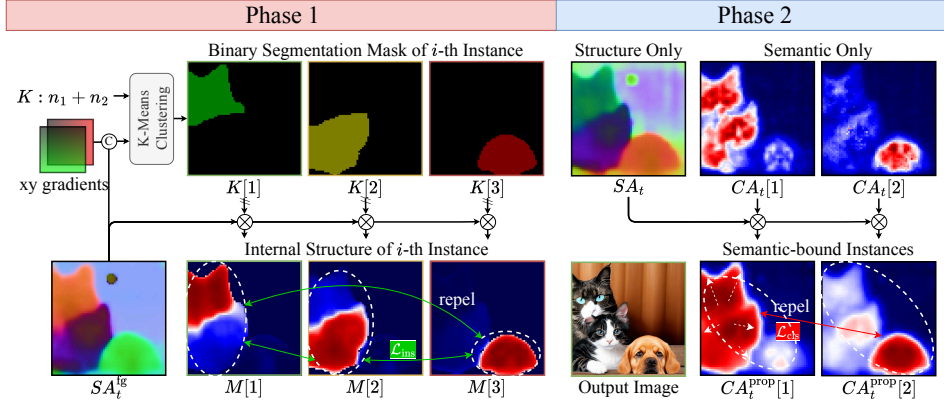


Figure 3: Overview of two phase control objective, ISAC. At phase 1, only the object counts work as a signal to separate instances, forming an objective \mathcal{L}_{ins} . At phase 2, where instance structures are formed, semantic signals are bind to instance structures. Now the phase 2 objective \mathcal{L}_{cls} ensures each object semantic corresponds to the most probable instances.

global foreground mask by self-to-cross class propagation¹ (Equation 3) followed by column-wise adaptive binarization (Equation 4):

$$CA_t^{\text{prop}}(X_t, \mathcal{T}) \leftarrow SA_t \cdot CA_t \in [0, 1]^{HW \times L} \quad (3)$$

$$CA_t^{\text{bin}}(X_t, \mathcal{T}) \leftarrow \text{Binarize}(CA_t^{\text{prop}}), \quad (4)$$

$$\text{where } \text{Binarize}(CA_t^{\text{prop}})[i, j] = \mathbf{1}[CA_t^{\text{prop}}[i, j] > \mu_j] \in \{0, 1\}$$

where μ_j is the column-wise mean. The binarized map $CA_t^{\text{bin}} \in \{0, 1\}^{HW \times L}$ contains foreground masks for target tokens τ_1, \dots, τ_k as well as non-target tokens (e.g., "the", "and"). The *global foreground mask* is the union of target tokens' masks: $M_{\text{fg}} = \bigcup_{\tau[i] \in \{\tau_j\}_{j=1}^k} CA_t^{\text{bin}}[:, i] \in \{0, 1\}^{HW}$.

Achieving instance structures with K-means clustering. From the global foreground mask M_{fg} , we index the self-attention to obtain the *filtered self-attention map* SA_t^{fg} as in equation 6:

$$\mathcal{I} = \{i : M_{\text{fg}}[i] = 1\}, \quad SA_t \leftarrow \mathcal{H}^{\text{self}}(X_t, \epsilon_\theta, t) \quad (5)$$

$$SA_t^{\text{fg}} \leftarrow SA_t[\mathcal{I}, \mathcal{I}] \in [0, 1]^{F \times F}, \text{ where } F := |\mathcal{I}| \quad (6)$$

For robust clustering, we concatenate to each row of SA_t^{fg} the corresponding normalized image coordinates $(x, y) \in [-1, 1]^2$ (one scalar per axis). We then apply K-means with $K = N$ to these augmented vectors, producing a one-hot assignment matrix $K \in \{0, 1\}^{F \times N}$ (Figure 4).

This yields cluster-wise structures via a single matrix product: $SA_t^{\text{fg}} K \in [0, 1]^{F \times N}$. The i -th column of $SA_t^{\text{fg}} K$ aggregates dependencies of pixels assigned to cluster i , and we treat it as a soft instance mask $M[i]$, can be interpreted as the internal structure of the i -th instance.

Measuring separation with Maximum Pixel-wise Overlap (MPO). Let $A, B \in [0, 1]^F$ be soft masks over F pixels. We define

$$\text{MPO}(A, B) = \max_{p \in \{1, \dots, F\}} (A[p] \cdot B[p]) \quad (7)$$

¹Class propagation is commonly used in segmentation methods (Shen et al., 2024; Wang et al., 2025b; Kipf & Welling, 2016; Zhu & Koniusz, 2021). We follow Shen et al. (2024)'s implementation.

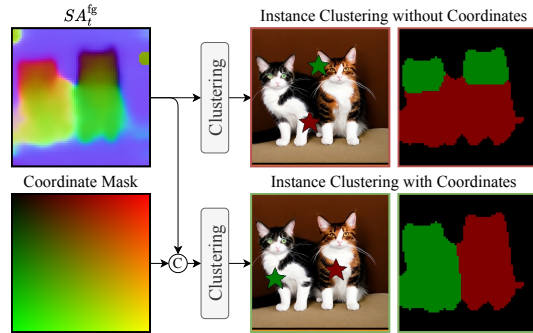


Figure 4: Adding X/Y coordinate masks improves instance clustering by introducing spatial cues, avoiding incorrect merging of separate instances.

which captures the peak local co-activation (worst-case overlap). Minimizing MPO suppresses even small but sharp collisions between masks, promoting spatial exclusivity that global similarities (e.g., IoU, KL) may under-penalize. We adopt MPO for both instance separation and semantic binding via \mathcal{L}_{ins} and \mathcal{L}_{cls} ; ablations and metric comparisons appear in §6.3 and Figure 6.

The Phase 1 objective, \mathcal{L}_{ins} , is therefore defined as the maximum MPO between any pair of instance structures:

$$\mathcal{L}_{\text{ins}}(X_t) = \max_{1 \leq i < j \leq N} \text{MPO}(M[i], M[j]) \quad (8)$$

5.2 PHASE 2: BINDING SEMANTICS TO INSTANCE STRUCTURES

Once instance structures stabilize, we bind class semantics to the formed regions following the rationale: “the semantic appearance of each instance should be bound to its corresponding instance structure.” We obtain semantic-bound maps by masking cross-attention with the formed structures (Equation 3), yielding $CA_t^{\text{cls}} \in [0, 1]^{H \times W \times k}$ for classes τ_1, \dots, τ_k . Let $\text{sign}(\tau_i, \tau_j) \in \{+1, -1\}$ encode the desired interaction between tokens: $+1$ for pairs that should *not* co-refer (different classes), -1 for pairs that *should* co-refer (attribute–object). Our Phase-2 objective is

$$\mathcal{L}_{\text{cls}}(X_t) = \max_{1 \leq i < j \leq k} \text{sign}(\tau_i, \tau_j) \cdot \text{MPO}(CA_t^{\text{cls}}[i], CA_t^{\text{cls}}[j]), \quad (9)$$

which penalizes peak overlap across different classes while encouraging peak co-activation for intended attribute binding. The relationship between tokens— $\text{sign}(\tau_i, \tau_j) \forall i, j$ —can be easily achieved with modern large language models (LLMs) as a syntax parser.

5.3 ALGORITHMIC APPLICATIONS

The overall two-phase control objective can be written in a generalized form:

$$\mathcal{L}_t(X_t) := \lambda_{\text{ins}}(t)\mathcal{L}_{\text{ins}}(X_t) + \lambda_{\text{cls}}(t)\mathcal{L}_{\text{cls}}(X_t) \quad (10)$$

If we denote the phase transition timestep as t^* , marking when instance formation is complete, the objective scheduling for this *hard* transition is defined with Heaviside step function H :

$$\lambda_{\text{ins}}(t) = 1 - H(t - t^*), \quad \lambda_{\text{cls}}(t) = H(t - t^*) \quad (11)$$

As the choice of schedule is flexible, we also propose a *soft* phase transition schedule, such as $\lambda_{\text{ins}}(t) = t/T$, $\lambda_{\text{cls}}(t) = 1 - t/T$, which eliminates the need to define t^* in advance. In our experiments, we empirically find that this *soft* schedule consistently outperforms the *hard* schedule from equation 11 when using $t^* = T/2$.

The ISAC objective can be directly injected to guide the generation of multiple instances in two ways. First, similar to previous methods (Chefer et al., 2023; Guo et al., 2024; Meral et al., 2024), by treating the objective as a loss function, we implement a latent optimization algorithm (Algorithm 1). Second, from another perspective, the objective can serve as a verifier to score the denoising trajectories of independent samples. This enables the selection of best output, analogous to inference-time scaling in SANA 1.5 (Xie et al., 2025), for which we present a latent selection algorithm in Algorithm 2.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Evaluation Metrics. We evaluate on T2I-CompBench (Huang et al., 2025) and HRS-Bench (Bakr et al., 2023), and additionally introduce a benchmark targeting similar-object scenarios. From T2I-CompBench we use the *color*, *texture*, and *complex* tasks to assess attribute binding; from HRS-Bench we use *spatial*, *size*, and *color* for attribute binding and instance-structure formation.

Our benchmark contains two settings: *multi-class* (multiple classes, one instance per class) and *multi-instance* (multiple instances of a single class). For *multi-class accuracy* (%), a class counts as correct if the generated instance matches the target class. For *multi-instance accuracy* (%), an instance counts as correct if it matches the target class and is spatially separated from other instances. Class tags in our benchmark prompts are drawn from COCO (Lin et al., 2014); prompt formatting appears in the Appendix. For multi-class accuracy, we enumerate all non-overlapping class pairs and report the mean. Detection uses Grounding-DINO (Liu et al., 2024) ensembled with YOLOv12 (Tian

Algorithm 1: Latent Optim. with ISAC

Input: Prompt \mathcal{T} , Model ϵ_θ , decoder \mathcal{D} ,
Learning rate η
Output: Image I_0 with multiple instances

```

1  $X_T \sim \mathcal{N}(0, I)$ 
2 for  $t = T, T-1, \dots, 1$  do
3    $\text{Denoise}(X_t, \mathcal{T}, \epsilon_\theta, t)$  with  $\mathcal{H}^{\text{self}}, \mathcal{H}^{\text{cross}}$ 
4    $SA_t \leftarrow \mathcal{H}^{\text{self}}(X_t, \mathcal{T}, \epsilon_\theta, t)$ 
5    $CA_t \leftarrow \mathcal{H}^{\text{cross}}(X_t, \mathcal{T}, \epsilon_\theta, t)$ 
6    $CA_t^{\text{prop}} \leftarrow SA_t \cdot CA_t$ 
7   Compute  $\mathcal{L}_{\text{ins}}, \mathcal{L}_{\text{cls}} \leftarrow \text{Eq. 4, 6, 8, 9}$ 
8    $\mathcal{L}_t(X_t) \leftarrow \lambda_{\text{ins}}(t)\mathcal{L}_{\text{ins}}(X_t) + \lambda_{\text{cls}}(t)\mathcal{L}_{\text{cls}}(X_t)$ 
9    $\tilde{X}_t \leftarrow X_t - \eta \cdot \nabla_{X_t} \mathcal{L}_t(X_t)$ 
10   $X_{t-1} \leftarrow \text{Denoise}(X_t, \mathcal{T}, \epsilon_\theta, t)$ 
11  $I_0 \leftarrow \mathcal{D}(X_0)$  // Decode to pixel

```

Algorithm 2: Latent Selection with ISAC

Input: Prompt \mathcal{T} , Model ϵ_θ , decoder \mathcal{D} , Batch size N
Output: Image I_0 with multiple instances

```

1  $X_T^{(i)} \sim \mathcal{N}(0, I), S[i] = 0, \forall i = 1, \dots, N$ 
2 for  $i = 1, \dots, N$  do
3   for  $t = T, T-1, \dots, 1$  do
4      $X_{t-1}^{(i)} \leftarrow \text{Denoise}(X_t, \mathcal{T}, \epsilon_\theta, t)$  with  $\mathcal{H}^{\text{self}}, \mathcal{H}^{\text{cross}}$ 
5      $SA_t \leftarrow \mathcal{H}^{\text{self}}(X_t^{(i)}, \mathcal{T}, \epsilon_\theta, t)$ 
6      $CA_t \leftarrow \mathcal{H}^{\text{cross}}(X_t^{(i)}, \mathcal{T}, \epsilon_\theta, t)$ 
7      $CA_t^{\text{prop}} \leftarrow SA_t \cdot CA_t$ 
8     Compute  $\mathcal{L}_{\text{ins}}, \mathcal{L}_{\text{cls}} \leftarrow \text{Eq. 4, 6, 8, 9}$ 
9      $\mathcal{L}_t(X_t^{(i)}) \leftarrow \lambda_{\text{ins}}(t)\mathcal{L}_{\text{ins}}(X_t^{(i)}) + \lambda_{\text{cls}}(t)\mathcal{L}_{\text{cls}}(X_t^{(i)})$ 
10    Score Update:  $S[i] \leftarrow S[i] + \mathcal{L}_t(X_t^{(i)})$ 
11  $i^* = \arg \min_i S[i]$  // Best scored latent
12  $I_0 \leftarrow \mathcal{D}(X_0^{(i^*)})$  // Decode to pixel

```

Table 1: Quantitative comparison of ISAC (Ours) and baseline methods on HRS Benchmark, T2I-CompBench and similar-object benchmark. All tasks handle multi-class scenarios.

Method	HRSBench			T2I-CompBench			Multi-Class Accuracy (\uparrow)				
	Color \uparrow	Spatial \uparrow	Size \uparrow	Color \uparrow	Texture \uparrow	Complex \uparrow	#2	#3	#4	#5	Average
SD1.5 (Rombach et al., 2022)	0.136	0.094	0.091	0.356	0.406	0.306	28%	2%	1%	0%	8%
+ A&E (Chefer et al., 2023)	0.149	0.104	0.101	0.392	0.447	0.290	48%	10%	5%	2%	16%
+ SynGen (Rassin et al., 2024)	0.159	0.111	0.107	0.420	0.479	0.311	50%	9%	4%	2%	16%
+ InitNO (Guo et al., 2024)	0.175	0.120	0.116	0.456	0.520	0.338	55%	12%	7%	5%	20%
+ TEBOpt (Chen et al., 2024a)	0.181	0.127	0.123	0.461	0.544	0.353	52%	11%	8%	3%	18%
+ ISAC (Ours)	0.318	0.263	0.252	0.683	0.631	0.354	65%	31%	29%	18%	36%
SD3.5-M (Esser et al., 2024)	0.425	0.264	0.209	0.796	0.726	0.377	62%	23%	12%	3%	25%
+ A&E (Chefer et al., 2023)	0.427	0.263	0.215	0.798	0.726	0.378	65%	29%	16%	5%	28%
+ SynGen (Rassin et al., 2024)	0.425	0.260	0.211	0.801	0.718	0.365	66%	28%	15%	6%	28%
+ InitNO (Guo et al., 2024)	0.443	0.275	0.228	0.810	0.728	0.378	77%	31%	17%	7%	33%
+ TEBOpt (Chen et al., 2024a)	0.438	0.279	0.220	0.805	0.730	0.381	78%	31%	19%	8%	34%
+ ISAC (Ours)	0.473	0.350	0.258	0.838	0.739	0.388	98%	51%	40%	20%	52%

et al., 2025) and YOLOE (Wang et al., 2025a) to reduce single-model errors; details of the ensemble process are in the Appendix.

Although ISAC addresses both settings, most baselines do not directly target the multi-instance task. To isolate the effect of diffusion dynamics, our main comparisons therefore focus on the multi-class setting; the multi-instance capability of ISAC is analyzed separately in Section F.

Implementation Details. We follow each method’s official guidance (CFG and sampling steps) for SD1.5 (Rombach et al., 2022) and SD3.5-M (Esser et al., 2024). Following the baseline methods, we apply latent optimization (Algorithm 1) with ISAC. Here, the only tunable hyperparameter is the learning rate η , which is fixed to 0.01 across models. Related ablation studies are included in the Appendix. Class tag-count pairs (τ_i, n_i) are extracted from prompts using an LLM parser; in multi-class settings $n_i = 1$ for all i , for all prompts.

6.2 MAIN RESULTS

Table 1 reports results for SD1.5 and SD3.5-M backbones, which represent UNet and DiT based diffusion models respectively. Across both models, ISAC achieves the best scores on HRS-Bench (Bakr et al., 2023), T2I-CompBench (Huang et al., 2025), and on the Multi-Class Accuracy metric, consistently outperforming all training-free baselines. These gains indicate ISAC improves not only instance-structure formation (*spatial/size/multi-class*) but also its subsequent attribute binding

Table 2: Effect of objective scheduling to multi-instance generation performance. Here, we use $\lambda_{\text{cls}}(t) = 1 - \lambda_{\text{ins}}(t)$ for the purpose of balancing the two components in Eq. 10, SD1.5 (Rombach et al., 2022) is the backbone model.

Config.	Description	$\lambda_{\text{ins}}(t)$	$\lambda_{\text{cls}}(t)$	Multi-Class	Multi-Instance
A	Only Instance Optimization	1	0	10% (-26%pt)	65% (-4 %pt)
B	Only Semantic Optimization	0	1	28% (-8 %pt)	54% (-15%pt)
C	Fixed Balance	0.5	0.5	25% (-11%pt)	60% (-9 %pt)
D	Semantic-to-Instance, Hard	$H(t - T/2)$	$1 - H(t - T/2)$	19% (-17 %pt)	52% (-17 %pt)
E	Semantic-to-Instance, Soft	$1 - t/T$	t/T	21% (-15 %pt)	55% (-14%pt)
F	Instance-to-Semantic, Hard	$1 - H(t - T/2)$	$H(t - T/2)$	35% (-1 %pt)	67% (-2 %pt)
G	Instance-to-Semantic, Soft (Ours)	t/T	$1 - t/T$	36%	69%

Table 3: Alternative similarity metrics for the proposed MPO in Eq. 7. We use soft transition schedule for all configurations.

Loss type	Multi-Class	Multi-Instance
MAE	9 % (-27%pt)	55% (-14%pt)
KL	16% (-20%pt)	60% (-9%pt)
IoU	20% (-16%pt)	61% (-8%pt)
MPO (Ours)	36%	69%

(*color/texture/complex*). Additional quantitative results in Table 13 shows consistent and model-agnostic gains with ISAC latent optimization algorithm.

Qualitative results in Figure 5 shows that in most cases SD1.5 (Rombach et al., 2022), A&E (Chefer et al., 2023), InitNO (Guo et al., 2024) (1st, 3rd rows) and SynGen (Rassin et al., 2024) (1st, 2nd row) fail to generate all 3 instances. Even succeed in generating 3 instances, they often wrongly allocate the class labels—A&E (Chefer et al., 2023) and InitNO (Guo et al., 2024) (2nd row) and SynGen (Rassin et al., 2024) (3rd row)—leading to another instance missing problem. Also in the case of SD1.5 (Rombach et al., 2022), A&E (Chefer et al., 2023) and InitNO (Guo et al., 2024) (2nd row), the left dog with brown hair indicates semantic features from adjacent instances flooded to it, which is a common failure mode of instance merging. In contrast, ISAC successfully generates 3 decoupled instances with distinct appearances, demonstrating its effectiveness in multi-instance generation.



Figure 5: Qualitative comparison using SD1.5 (Rombach et al., 2022) as a backbone and added attention control methods. For all cases, “A photo of two cats and a dog” is an input prompt.

6.3 ABLATION STUDY

We ablate the fundamental effect of our method design to instance structure formation and instance-semantic binding. Since our similar-object benchmark handles that basic scenarios and as it supports instance count-wise report, we majorly use it for ablation study and discussions.

Contribution of Loss Components. We ablate the instance-decoupling loss \mathcal{L}_{ins} and instance-semantic binding loss \mathcal{L}_{cls} by selectively disabling each term (Table. 2). With SD1.5, optimizing *only* \mathcal{L}_{ins} (Config. A) yields 10% multi-class and 65% multi-instance accuracy; optimizing *only* \mathcal{L}_{cls} (Config. B) gives 28% and 54%; using constant weights (Config. C) gives 25% and 60%. Both components are therefore necessary for balanced performance.

Scheduling instance-to-class dynamics. We compare weighting schedules for $(\lambda_{\text{ins}}(t), \lambda_{\text{cls}}(t))$ in Table. 3. The *soft two-phase* schedule (Ours; Config. F) attains **36%** multi-class and **69%** multi-instance accuracy, outperforming a *reversed* class-to-instance schedule (Config. D), which underperforms even the constant baseline (Config. C). This supports aligning optimization with diffusion dynamics: prioritize instance separation early, then refine semantic assignment. Also, *soft*

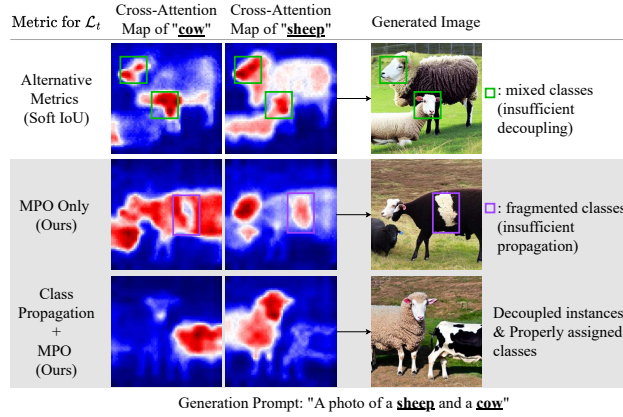


Figure 6: Effect of MPO and Class Propagation. Compared to the baseline metric, our full method produces better instance decoupling and class consistency in multi-class prompts.

two-phase marginally outperforms the *hard* transition (Config. E), verifying the soft two-phase design is valid for temporal dynamics of diffusion models.

Effectiveness of Maximum Pixel-wise Overlap (MPO) and Class Propagation. Replacing MPO with alternative similarities substantially degrades accuracy (Table 3). With MPO, multi-class and multi-instance accuracies reach **36%** and **69%**, respectively, versus **20%/61%** for IoU, **16%/60%** for KL, and **9%/55%** for MAE. These results indicate that MPO’s peak-overlap penalty enforces strict local exclusivity, improving both instance separation and semantic binding.

Figure 6 further illustrates these effects. Optimizing with global metrics (e.g., Soft IoU) yields overlapping cross-attention for different classes (“cow”/“sheep”), producing mixed activations and semantic leakage in the image. Using MPO alone removes most overlaps but can leave fragmented, under-propagated class activations. Combining MPO with class propagation resolves both issues, yielding well-localized attention maps and cleanly separated classes in the final image.

7 CONCLUSION

We presented ISAC, a training-free objective that explicitly separates instance formation from semantic binding, aligning inference control with the dynamics of diffusion models. By enforcing early structural exclusivity and subsequent semantic alignment, ISAC overcomes the common failure modes of instance merging and omission. Across diverse benchmarks and backbones, ISAC consistently improves multi-instance fidelity and integrates seamlessly with layout-guided or fine-tuned pipelines. These results suggest that instance-first dynamics are a fundamental principle for multi-object generation, opening new directions for extending control to video, medical imaging, and distilled diffusion models.

REFERENCES

- Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20041–20053, 2023.
- Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.
- Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

- Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. *arXiv preprint arXiv:2406.10210*, 2024.
- 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, 2023.
- Chieh-Yun Chen, Chiang Tseng, Li-Wu Tsao, and Hong-Han Shuai. A cat is a cat (not a dog!): Unraveling information mix-ups in text-to-image encoders through causal analysis and embedding optimization. *Advances in Neural Information Processing Systems*, 2024a.
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023a.
- Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024b.
- Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023b.
- Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation, 2024. URL <https://arxiv.org/abs/2410.14324>.
- Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation, 2024.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=PUIqjT4rzq7>.
- Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*, 2024.
- 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.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Michael Hobley and Victor Prisacariu. Learning to count anything: Reference-less class-agnostic counting with weak supervision. *arXiv preprint arXiv:2205.10203*, 2022.
- Matthew Honnibal. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. (*No Title*), 2017.
- Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. *Advances in Neural Information Processing Systems*, 37: 137646–137672, 2024.

- Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025. doi: 10.1109/TPAMI.2025.3531907.
- Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024.
- Wonjun Kang, Kevin Galim, Hyung Il Koo, and Nam Ik Cho. Counting guidance for high fidelity text-to-image synthesis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 899–908. IEEE, 2025.
- Kwanyoung Kim and Jong Chul Ye. Noise2score: tweedie’s approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021.
- Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. In <https://arxiv.org/abs/2410.05591>, 2024.
- Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- Phillip Y. Lee, Taehoon Yoon, and Minhyuk Sung. Groundit: Grounding diffusion transformers via noisy patch transplantation. In *Advances in Neural Information Processing Systems*, 2024.
- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023.
- Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9005–9014, 2024.
- Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. *The Thirteenth International Conference on Learning Representations*, 2025.
- Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention re-focusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7932–7942, 2024.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

- Weimin Qiu, Jieke Wang, and Meng Tang. Self-cross diffusion guidance for text-to-image synthesis of similar subjects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23528–23538, 2025.
- 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, 2024.
- Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.
- Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9370–9379, 2024.
- Takahiro Shirakawa and Seiichi Uchida. Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.
- Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything, 2025a. URL <https://arxiv.org/abs/2503.07465>.
- Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *IEEE Transactions on Image Processing*, 2025b.
- Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation, 2024a.
- Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Text-to-image diffusion with token-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8553–8564, June 2024b.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL <https://arxiv.org/abs/2508.02324>.
- Jiayu Xiao, Liang Li, Henglei Lv, Shuhui Wang, and Qingming Huang. R&b: Region and boundary aware zero-shot grounded text-to-image generation, 2023.
- Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng YU, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=27hOkXzy9e>.

- Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7452–7461, 2023.
- Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *International Conference on Machine Learning*, 2024.
- Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kaini Wang, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 2024a.
- Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. *arXiv preprint arXiv:2410.07171*, 2024b.
- Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6818–6828, 2024a.
- Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. *arXiv preprint arXiv:2410.12669*, 2024b.
- Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International conference on learning representations*, 2021.

APPENDIX

A USAGE OF LARGE LANGUAGE MODELS.

A large language model (LLM) was used solely for light editorial assistance—e.g., correcting typographical errors and refining phrasing for academic style. The LLM did not contribute to research conception, experimental design, implementation, or analysis. All scientific content and conclusions are the authors’ own, and the authors bear full responsibility for the manuscript.

B LIMITATIONS

Latency and VRAM Overhead. While ISAC introduces an increase in latency and VRAM usage—typically around $2\times$ to $3.3\times$ compared to base models (see Table 1)—this overhead is a known limitation inherent to latent optimization-based methods. Importantly, the computational cost remains comparable to existing approaches like A&E, SynGen, and InitNO. Moreover, the overhead does not scale significantly with the number of target instances or classes, demonstrating ISAC’s efficiency and robustness in complex settings.

C IMPLEMENTATION DETAILS

C.1 METHOD DETAILS

Choice of the Learning Rate η . We conducted experiments to evaluate the sensitivity of our method to the learning rate η (Algorithm 1 with our time-dependent coefficients $\lambda_{\text{ins}}(t) = 1 - t/T$ and $\lambda_{\text{cls}}(t) = t/T$). When the learning rate is too low, the optimization loss does not converge, and thus has little to no effect on the outcome. Conversely, excessively high learning rates lead to latent collapse, resulting in image quality degradation. Existing studies that use latent optimization methods all share this same limitation, and our method is no exception.

We found that a learning rate of $1e-2$ is sufficient for stable optimization. Additionally, we tested values in the range of $5e-3$ to $1e-1$ and observed consistent trends in Figure 7. We clarify that the gradient descent step $\tilde{X}_t \leftarrow X_t - \eta \nabla_{X_t} \mathcal{L}_t(X_t)$ in Algorithm 1 is taken once per timestep. η is the only hyperparameter in the entire ISAC pipeline.

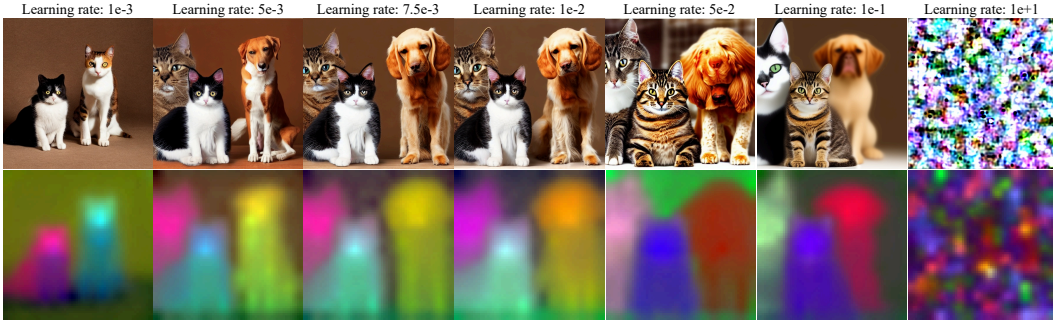


Figure 7: Qualitative comparison of ISAC with various learning rates. For all cases, we provide “A photo of two cats and a dog” as the input prompt and use SD1.5 (Rombach et al., 2022) as the baseline diffusion model.

Choice of Clustering Algorithm. We explored alternative clustering algorithms in Table 4 and Figure 8. While all three (K-means, Spectral Clustering, Gaussian Mixture Model) algorithms exhibit similar performance in terms of both multi-class and multi-instance accuracy, K-means (our choice) is $\times 3$ to $\times 16$ faster.

Attention Accumulation. We accumulate self- and cross-attention maps with all spatial resolutions for each denoising step. Then, we apply min-max normalization to the accumulated maps as shown

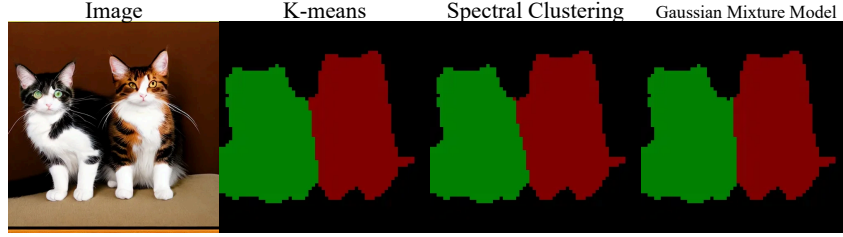


Figure 8: Qualitative comparison of clustering algorithms. The image is generated with the prompt, "a photo of two cats". This figure is an extension of Figure 4.

Table 4: Comparison of clustering algorithms in terms of accuracy (%) and latency (ms). Latency is defined as the execution time for a one-time application of each clustering algorithm to a self-attention map. This table can be seen as an extension of Table 2.

Clustering Algorithm	Multi-class Accuracy (\uparrow)	Multi-Instance Accuracy (\uparrow)	Latency (ms) (\downarrow)
K-means Clustering	36%	69%	505
Spectral Clustering	35%	69%	1,630
Gaussian Mixture Model	36%	69%	8,313

in Figure 9. For fair comparisons, we use the same attention accumulation scheme for all baseline methods. Details on attention maps and accumulation are given in Section 3.

C.2 BASELINE IMPLEMENTATION DETAILS

Cross-attention Normalization. Baseline methods, Attend-and-Excite (Chefer et al., 2023), InitNO (Guo et al., 2024), and TEBOpt (Chen et al., 2024a), utilized the the `softmax` normalization technique on cross-attention maps. It operates by applying the `softmax` function to the cross-attention maps along the token dimension, excluding the `SOT` token at index 0. This sharpens the attention, emphasizing foreground objects while suppressing background noise. The formulation is given by:

$$CA_t^{\text{softmax}} = \text{softmax}(\tau \cdot CA_t[1:]). \quad (12)$$

Following the baseline implementation, we set $\tau = 100$ for SD1.5 (Rombach et al., 2022). SD1.4 and SD2.1 models (Rombach et al., 2022) that share the same architecture as SD1.5 are also tested with the same temperature values in Section G.

We found that the temperature hyperparameter τ plays a critical role in performance. A large τ produces overly sharp cross-attention maps, causing the model to over-focus on a single token and potentially overlook relevant context. In contrast, a small τ results in overly smooth attention maps, which can dilute the focus on foreground objects. As shown in Figure 9, directly applying the temperature value from SD1.5 (Rombach et al., 2022) to SDXL (Podell et al., 2023) or SD3.5-M (Esser et al., 2024) leads to undesirable behaviors. For example, in SDXL, a non-instance token such as "of" becomes the most attended token, while in SD3.5-M, the signal is overly diffused after the `softmax` normalization.

Though the performance is sensitive to τ , along with the lack of official implementations for SD3.5-M (Esser et al., 2024), we adopt $\tau = 100$ and evaluate baseline methods to fairly compare the performance on Diffusion Transformer based diffusion models.

Min-max Normalization in ISAC. In contrast, ISAC requires no temperature tuning. Instead, it adopts a simple element-wise min-max normalization to rescale attention maps:

$$CA_t^{\text{minmax}} = \frac{CA_t - \min(CA_t)}{\max(CA_t) - \min(CA_t)}. \quad (13)$$

As illustrated in Figure 9, while min-max normalization may retain some background noise, it preserves object-relevant signals more reliably across models. This modification was essential for

extending our method to other architectures such as SDXL (Podell et al., 2023), SD3.5-M (Esser et al., 2024), and PixArt- α (Chen et al., 2023a). However, since it is a relatively minor adjustment, we discuss it in the appendix rather than in the main text.

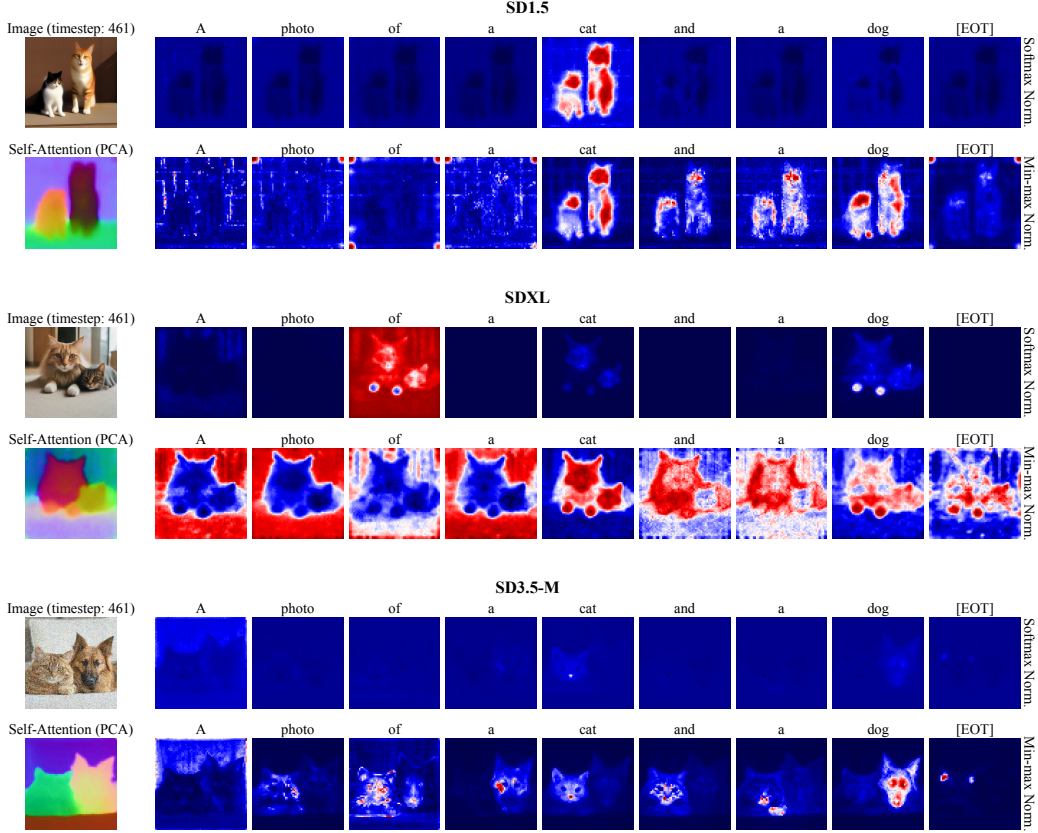


Figure 9: Comparison of cross attention maps after applying softmax and min-max normalization on SD1.5 (Rombach et al., 2022), SDXL Podell et al. (2023), SD3.5-M (Esser et al., 2024). Here, we used a fixed temperature hyperparameter $\tau = 100$ across all models.

C.3 OUR BENCHMARK: MULTI-CLASS AND MULTI-INSTANCE ACCURACY

Inspired by the MultiGen benchmark (Wang et al., 2024b), we evaluate ISAC’s ability to generate (1) multiple instances of different classes (multi-class accuracy) and (2) multiple instances of the same class (multi-instance accuracy). Notably, our multi-class evaluation is designed to be more challenging than the original MultiGen setup.

From our observations in Figure 10, text-to-image diffusion models tend to struggle more when generating multiple intra-category instances (e.g., dog and cat) compared to inter-category instances (e.g., animal and vehicle). Based on this insight, our multi-class benchmark emphasizes scenarios where all classes are sampled from the same semantic category.

To facilitate quantitative evaluation, we use a subset of countable object classes from the 80 COCO categories² (Lin et al., 2014), grouped into four higher-level categories: animals, vehicles, sports, and food. The full list of classes and their categorical assignments is provided in Table 5.

We excluded certain COCO classes from our evaluation for the following reasons: The person class was removed due to the difficulty of instance differentiation caused by high variability in pose and

²Most detection models are trained on COCO classes, so we focus our evaluation on COCO, where the pretrained models are most reliable, rather than experimenting with new classes from datasets like ADE. In practice, benchmarks such as TokenCompose have shown minimal differences in performance trends between COCO and ADE.

(a) A photo of a cat and a bicycle



(b) A photo of a cat and a dog



Figure 10: Inter-category and intra-category instance generation with 10 random seeds using SDXL (Podell et al., 2023). (a) Inter-category instance generation, (b) Intra-category instance generation. We can see that both instances appear in inter-category images with some quality loss (8/10), however, only 3/10 can correctly generate each instance in intra-category images.

Table 5: Countable classes from COCO (Lin et al., 2014) dataset used in the evaluation.

Category	Classes
Animal	(9 classes) cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe
Vehicle	(8 classes) bicycle, car, motorcycle, airplane, bus, train, truck, boat
Sports	(10 classes) skateboard, snowboard, skis, sports ball, baseball bat, baseball glove, tennis racket, surfboard, kite, frisbee
Food	(10 classes) banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake

posture. We also excluded small or background-like objects such as fork, spoon, knife, keyboard, remote, and toothbrush, which are challenging to detect reliably. In addition, we omitted scene or ambiguous objects like tv, book, clock, vase, bed, and couch, where instance boundaries are often unclear. Finally, we removed objects such as bench, bowl, chair, and dining table, where duplication or spatial overlap lack meaningful semantic distinction.

Multi-class Accuracy. For each evaluation, we randomly sample k classes ($2 \leq k \leq 5$) from a single category and compose a prompt in the form: “A photo of a [class A], a [class B], ..., and a [class E].” For example, when sampling from the animal category, the prompt could be: “A photo of a dog, a cat, a horse, a cow, and a sheep.” The image is then generated using a text-to-image diffusion model conditioned on this prompt.

Multi-instance Accuracy. In this setup, we randomly sample a single class A and specify the number of instances n ($2 \leq n \leq 5$) to generate. The corresponding prompt is structured as: “A photo of [n] [class A]s.” For instance, if the selected class is “cat” and $n = 5$, the prompt would be: “A photo of five cats.” The model is then evaluated on its ability to correctly generate the specified number of instances of the given class.

Properties	Attention Control Methods	Model-Aided Methods	Ours
Semantic decoupled instance formation first approach	✗	✗	✓
Can separate instances with semantic control	✓	✓	✓
Can separate instances with only structural info.	✗	✓	✓
Do not require external vision model	✓	✗	✓

Table 7: Comparison of capabilities and setups across related methods.

Evaluation via Ensemble. To evaluate the accuracy, we use an ensemble of three state-of-the-art detection models to reduce single detector errors (open-vocabulary detectors—Grounding DINO (Liu et al., 2024) and YOLOE (Wang et al., 2025a) & closed-vocabulary detector trained on COCO (Lin et al., 2014)—YOLOv12 (Tian et al., 2025)). Given a generated image, each model detects the instances in the image. We only leave the detected instances where each instance is captured by any two detectors. The accuracy is then calculated as the ratio of the number of correctly detected instances to the total number of instances. For example, if “cat” and “dog” are detected in the image from the text prompt “A photo of a cat, a dog, a horse, a cow and a sheep”, then the accuracy is calculated as $2/5 = 40\%$ for multi-class accuracy. For multi-instance accuracy, if 3 instances of “cat” are detected in the image from the text prompt “A photo of five cats”, then the accuracy is calculated as $3/5 = 60\%$.

The choice of text prompts and the number of images to be generated are determined as follows. For **multi-class evaluation**, we can achieve all possible combinations of classes to build the text prompts. The number of combinations for each category is shown in Table 6. We randomly sample 20% of the combinations for each category and generate 10 images for each text prompt. For example, in #5-class evaluation, we randomly sample 25 combinations for animal category, 11 combinations for vehicle category, 50 combinations for sports and food categories respectively. In total, we generate 10 images for each of the 136 ($= 25 + 11 + 50 + 50$) text prompts. For **multi-instance evaluation**, we use each class in the category to build a single text prompt. Therefore, we have 9 text prompts for animal category, 8 text prompts for vehicle category, and 10 text prompts for sports and food categories respectively. We generate 10 images for each text prompt. In total, we generate 10 images for each of the 37 ($= 9 + 8 + 10 + 10$) text prompts.

Table 6: Possible combinations of countable classes for multi-class evaluation.

Category	#2	#3	#4	#5
Animal (9 classes)	36	84	126	126
Vehicle (8 classes)	28	56	70	56
Sports (10 classes)	45	120	210	252
Food (10 classes)	45	120	210	252
Total	154	380	616	686

D BROADER RELATED WORK COMPARISONS

To improve the generation of multi-instance images, many training-free methods have been proposed. However, focusing on semantic-level guidance (Chefer et al., 2023; Chen et al., 2024a; Feng et al., 2023; Rassin et al., 2024; Guo et al., 2024; Hu et al., 2024; Meral et al., 2024; Kwon & Ye, 2024; Shen et al., 2024) is not sufficient to control the instance formation. To overcome this limitation with spatial information, some methods introduce additional instance annotations such as bounding box (Bar-Tal et al., 2023; Shirakawa & Uchida, 2024; Xie et al., 2023; Chen et al., 2023b; Xiao et al., 2023; Dahary et al., 2024; Lee et al., 2024; Lian et al., 2023; Park et al., 2025) or exploit external models’ prior (Kang et al., 2025). In spite of these strong supervision over instances, they still fail to form correct instance structures since they overlooked the importance of forming instance structure in early diffusion steps.

In contrast to these methods, our proposed ISAC focus on instance formation in early diffusion steps. We achieve this by utilizing a hierarchical, tree-structured prompt mechanism that first determines the instance count and then adds semantic information. Since instance counts are easily parsed from text prompts, it adds minimal overhead to prompting compared to other methods that require additional annotations or external models. ?? summarizes the differences between ISAC and other methods.

Semantic-aware classifier-free guidance (S-CFG) (Shen et al., 2024) points out the spatial inconsistency of the global guidance method, Classifier-free guidance (CFG) (Ho & Salimans, 2022),

Table 8: Quantitative comparisons of ISAC with external model-aided methods. **Bold** indicates the best performance, underline indicates the second best performance.

Method	Use Supervision? (weak/strong)		Multi-Instance Accuracy (↑)					Latency (↓)	VRAM (↓)
	Instance Counts	External Model	#2	#3	#4	#5	Average		
SD1.4 (Rombach et al., 2022)	✗	✗	94%	74%	28%	22%	55%	8s	4.9GB
+ Counting Guidance (Kang et al., 2025)	✓	✓ (RCC (Hobley & Prisacariu, 2022))	79%	67%	32%	19%	49%	14s	17.5GB
+ ISAC (Ours)	✓	✗	100%	90%	51%	40%	70%	21s	9.7GB

and proposes a semantic region-level guidance method leveraging self- and cross-attention maps. However, S-CFG still lacks the ability to control the overall instance layout, as it only focuses on semantic-level guidance.

Multi-Concept Resampling, where we abbreviate it to MCR, proposed by (Kwon & Ye, 2024) is a method that mixing the noise predictions of single class with the noise predictions of multiple classes. It is aimed to preserve the semantic information of the single instance while generating multiple classes. However, lack of considering the spatial information of the instance, MCR is not able to control the instance formation.

Therefore, controlling the instance formation requires understanding on spatial positioning of instances. Some methods (Bar-Tal et al., 2023; Shirakawa & Uchida, 2024; Xie et al., 2023; Chen et al., 2023b; Xiao et al., 2023; Dahary et al., 2024; Lee et al., 2024; Lian et al., 2023; Park et al., 2025) utilize **bounding box layout** to provide where the instances should be placed. However, these methods still struggle with overlapping instances (Yang et al., 2024). Semantic driven guidance, which is equivalent to separating class appearances with cross-attention maps, is performed on incomplete instance formation with vague boundaries. Therefore, these methods often merge adjacent instance formations into a single, leading to a missing instance.

Another approach is to leverage pretrained vision model to aid the instance formation. **Counting Guidance** (Kang et al., 2025), where we abbreviate it to CG, proposes a classifier guidance method that leverages the pretrained vision model to aid the instance formation. Specifically, inspired by the Universal Guidance (Bansal et al., 2023), CG applies the counting network (Hobley & Prisacariu, 2022) (trained to count the number of instances in an image) to intermediate images denoised with Tweedie’s formula (Kim & Ye, 2021). Using the prediction of the counting network, CG then guides the diffusion process to generate the desired number of instances. However, blurry and semantically poor intermediate images in early diffusion steps make the pretrained vision model less effective. As illustrated in Figure 11, it is hard to tell what each instance is in the intermediate images and the state-of-the-art vision models, such as Grounding DINO (Liu et al., 2024) and YOLOE (Wang et al., 2025a), fail to provide accurate predictions. On the other hand, vision models are trained on clean, semantically rich images. This domain mismatch makes predictions of pretrained vision models in early diffusion steps less effective. Table 8 supports this observation, showing that the counting guidance (Kang et al., 2025) is less effective than ISAC in terms of instance formation.

Additionally, while universal guidance (Bansal et al., 2023) propose to apply vision model predictions on tweedie-denoised (Kim & Ye, 2021) images, tweedie formula cannot be applied to flow matching (Lipman et al., 2022) based models, such as SD3.5-M (Esser et al., 2024). This even hinders the application of the pretrained vision models to flow matching based models.

E DYNAMICS ANALYSIS OF DIFFUSION MODELS

E.1 SELF-ATTENTION MAPS IMPLIES INSTANCE FORMATION

In Figure 11, we visualize the evolution of denoising process. It contains noisy images, PCA visualization of self-attention maps, cross-attention maps on a token “cats” and detection models’ predictions on each noisy image. From the evolution of noisy images, we can see that instance structures are formed in the early denoising steps (until 3rd column), and the instance structures are preserved in the later denoising steps. Self-attention maps follow the same trend, where the attention maps are more focused on the instance structures in the early denoising steps. This shows that self-attention maps are able to capture the instance formation process of diffusion models. Predictions of detection models also share the same turning point that after the instance structures are formed

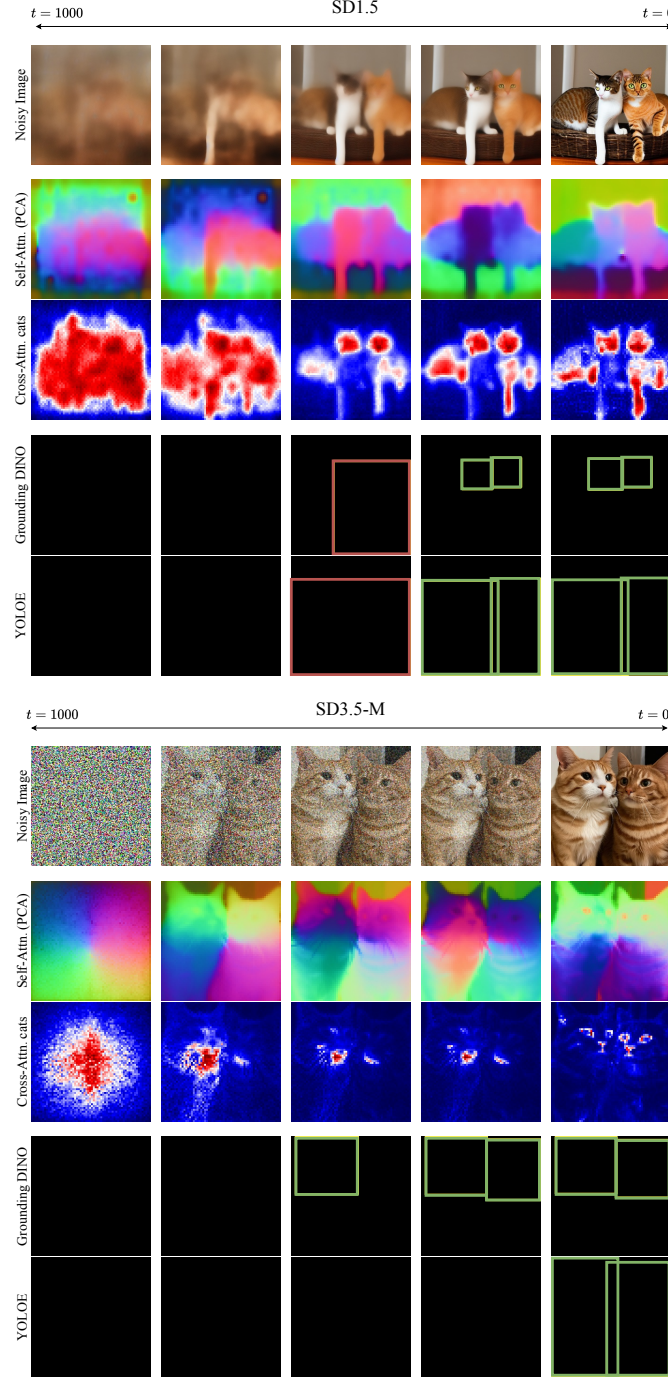


Figure 11: Dynamics of text-to-image diffusion models. For SD1.5 (Rombach et al., 2022), we visualize the tweedie-denoised (Kim & Ye, 2021) image as noisy images, then apply detection models on the denoised images following universal guidance (Bansal et al., 2023). Self-attention maps are visualized by PCA projection of the attention maps. The attention maps are obtained as explained in Section C. For each bounding box prediction of the detection model, if it is correct, we color it in green, and if it is incorrect, we color it in red.

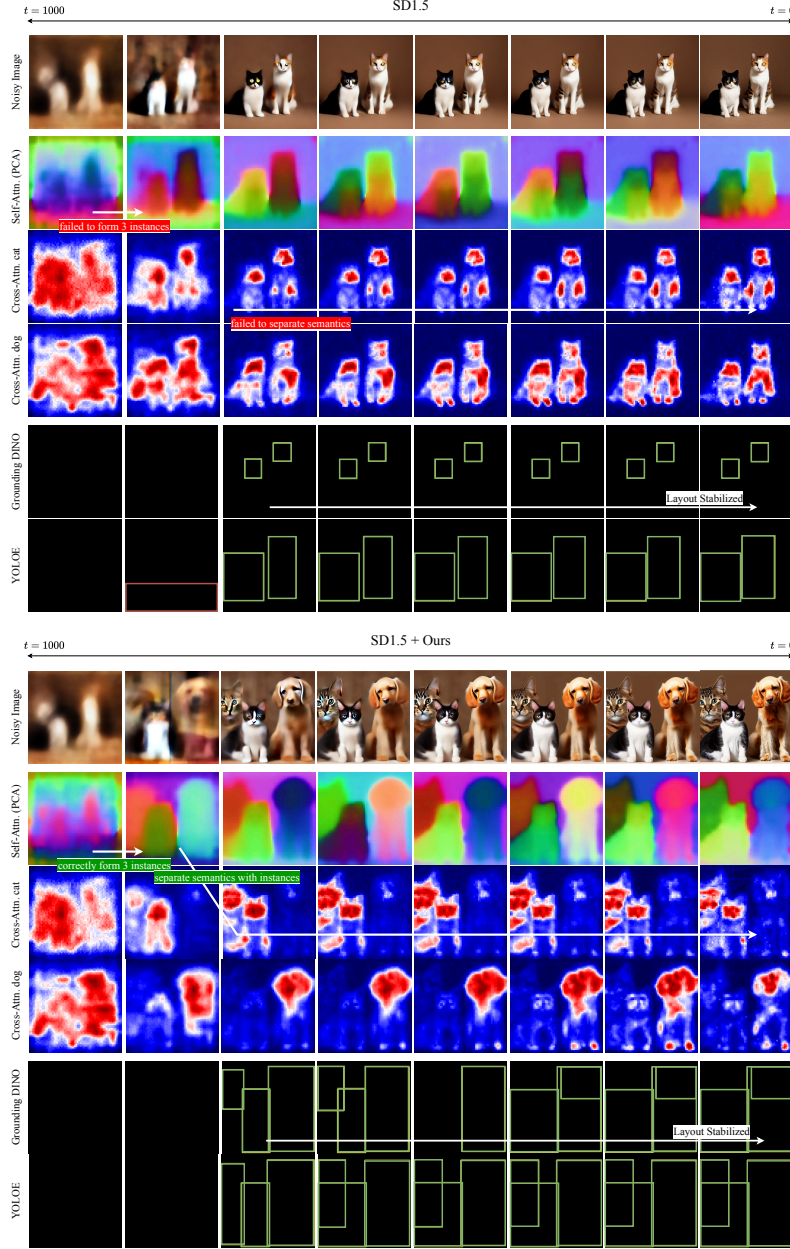


Figure 12: Qualitative comparisons of evolving dynamics of SD1.5 (Rombach et al., 2022) with and without ISAC control. We visualize the tweedie-denoised (Kim & Ye, 2021) image as noisy images, then apply detection models on the denoised images. Self-attention maps are visualized by PCA projection of the attention maps. The attention maps are obtained as explained in Section C. For each bounding box prediction of the detection model, if it is correct, we color it in green, and if it is incorrect, we color it in red.

and each instance become semantically identifiable, the predictions become correct. Semantically poor images in early denoising steps, where instance formation is actively done, make the use of pretrained vision models to aid instance formation process less effective.

E.2 VISUALIZATION OF DYNAMICS ALIGNED ATTENTION CONTROL

Figure 12 shows how ISAC’s instance forming first approach enhances multi-instance generation. ISAC first applies guidance to the self-attention maps, which helps to form correct number of instances in the early denoising steps. Then, leveraging the stand out instance structures, ISAC applies guidance to the instance-aware cross-attention maps to encourage each class appearance to be separated into its dedicated instance structure. Without instance control, however, SD1.5 (Rombach et al., 2022) not only fail to form correct number of instances, but also class semantics spread across the instance structures.

It is worth note that the instance forming guidance should be applied in the early denoising steps, otherwise the instance structures are already formed and the guidance will not be effective. The prediction results of Grounding DINO (Liu et al., 2024) and YOLOE (Wang et al., 2025a) also support that their predictions are only effective after each instance in noisy image contains distinguishable semantic information. This mismatch highlights ISAC’s advantage of instance forming first approach.

F DISCUSSION

ISAC’s Ability to Generate Multiple Instances from a Single Object Class. ISAC can control generation of more than one instance from a single object class. To evaluate its performance we evaluate ISAC on T2I-CompBench (Huang et al., 2025) *numeracy* task and *multi-instance* task in Table 9. The results show ISAC is also effective in multi-instance tasks, and notably SD1.5 (Rombach et al., 2022) with ISAC excels the plain SDXL (Podell et al., 2023)’s performance. Table 13 show the full results with multi-instance task, ISAC consistently outperforms across various diffusion models. Table 9: Quantitative results on T2I-CompBench Numeracy task and ours multi-instance accuracy task. This shows the performance on generating multiple instances from each object class. **Bold** indicates the best and underline indicates the second best performance.

Method	# Parameters	T2I-CompBench	Multi-Instance Accuracy (↑)				
		Numeracy (↑)	#2	#3	#4	#5	Average
SD1.5 (Rombach et al., 2022)	0.8B	46.5%	88%	65%	36%	26%	54%
+ ISAC (Ours)	0.8B	<u>54.3%</u>	95%	82%	56%	44%	69%
SDXL (Podell et al., 2023)	2.6B	50.7%	90%	71%	49%	32%	61%
+ ISAC (Ours)	2.6B	63.4%	96%	89%	71%	47%	76%

ISAC Extension with Fine-tuned Models. We apply ISAC on top of two finetuning approaches that enhance spatial understanding via external supervision: *TokenCompose* (Wang et al., 2024b) and *IterComp* (Zhang et al., 2024b). TokenCompose first generates an image, obtains segmentation maps using Grounded SAM (Ren et al., 2024). And fine-tunes the model to align cross-attention with those masks, improving multi-class generation. IterComp performs iterative, preference-guided refinement using a gallery of specialized models, leveraging RPG (Yang et al., 2024) and InstanceDiffusion (Wang et al., 2024a) to strengthen spatial reasoning. As shown in Table 10, ISAC further improves both multi-class and multi-instance accuracy when added to these baselines. Note that TokenCompose uses float32 weights, roughly doubling latency and VRAM compared to SD1.4.

Table 10: Extension of ISAC to fine-tuned models.

Method	Multi-Class Accuracy (↑)					Multi-Instance Accuracy (↑)					Latency (↓)	VRAM (↓)
	#2	#3	#4	#5	Average	#2	#3	#4	#5	Average		
TokenCompose _{SD1.4} (Wang et al., 2024b)	27%	4%	1%	0%	8%	77%	65%	41%	13%	49%	12s	8.5GB
+ ISAC (Ours)	62%	36%	28%	17%	36%	84%	80%	60%	30%	63%	33s	17.9GB
IterComp _{SDXL} (Zhang et al., 2024b)	11%	5%	4%	0%	5%	95%	73%	64%	37%	67%	49s	11.8GB
+ ISAC (Ours)	46%	28%	26%	21%	30%	99%	93%	85%	55%	83%	100s	29.9GB

ISAC Extension with Layout-to-Image Models. Table 11 show ISAC can easily be added on a layout-to-image model GLIGEN (Li et al., 2023), since ISAC only need to access latents and the

model’s attention layers. The compared method Attention-Refocus (Phung et al., 2024) is an attention driven latent optimization technique, forcing the cross/self attention maps of instances to attend only in predefined layouts (bounding-box inputs). ISAC, on the other hand, can achieve dense instance mask layouts without box layouts, shows its effectiveness even with layout-to-image models.

Table 11: ISAC extension to Layout-to-Image model, GLIGEN.

Method	Use layout in guidance?	HRSBench			
		Counting \uparrow	Color \uparrow	Spatial \uparrow	Size \uparrow
GLIGEN (Li et al., 2023)	-	66.58	30.74	26.75	18.78
+ Attention-Refocus (Phung et al., 2024)	✓	67.54	40.22	27.74	26.32
+ ISAC (Ours)	✗	71.28	45.21	28.12	27.51

Latent Selection with ISAC is Scalable. Latent optimization with ISAC requires huge VRAM usage due to its gradient computation through the model, typically around $2\times$ to $3.3\times$ compared to base models (Table 1). This restricts general applicability to huge models such as Flux (Labs, 2024) and Qwen-Image (Wu et al., 2025). Yet, latent selection with ISAC is comparably efficient, because no-gradient computation is required and image generation can be processed in a batch. Table 12 demonstrates that ISAC is even effective in modern diffusion models with comparably smaller overhead increase.

Table 12: Latent selection with ISAC. Best-of-1 selection is applied with 10 generated images.

Method	Multi-Class Accuracy (\uparrow)					Multi-Instance Accuracy (\uparrow)					Latency (\downarrow)	VRAM (\downarrow)
	#2	#3	#4	#5	Average	#2	#3	#4	#5	Average		
Flux.1-dev (Labs, 2024)	84%	37%	3%	2%	31%	97%	89%	82%	66%	83%	50s	37.2GB
+ ISAC (Ours)	97%	48%	38%	19%	51%	99%	94%	85%	72%	88%	85s	40.8GB
Qwen-Image (Wu et al., 2025)	91%	45%	33%	10%	48%	98%	92%	84%	70%	86%	140s	60.1GB
+ ISAC (Ours)	99%	58%	42%	25%	56%	99%	96%	89%	78%	91%	210s	65.3GB

G ADDITIONAL QUANTITATIVE RESULTS

We provide full quantitative results of similar object benchmark in Table 13. The results are obtained by applying ISAC to various models, including SD1.4, SD1.5, SD2.1 (Rombach et al., 2022), SDXL (Podell et al., 2023), SD3.5-M (Esser et al., 2024), PixArt- α (Chen et al., 2023a), and PixArt- Σ (Chen et al., 2024b).

H ADDITIONAL QUALITATIVE RESULTS

Generality Beyond Animals and Simple Prompts. Readers may be interested in whether our method is effective for more complex sentence structures and a wider variety of object classes. We provide qualitative results that include prompts with diverse object classes and relational structures in Figure 13. All prompts are taken from T2I-CompBench++ (Numeracy) (Huang et al., 2025) benchmark.

These results highlight the generality of ISAC to a wide range of object classes and complex prompts. Aligning with the quantitative gains across a wide range of classes and text prompt complexity, our method consistently improves the quality of generated images, demonstrating its robustness and versatility.

Robust Multi-instance Generation Across Seeds. We provide qualitative results when generating 5 images with a fixed prompt, "three cats", on SD1.5 Rombach et al. (2022) and SD3.5-M Esser et al. (2024). When the baseline output already contains the correct number of instances, ISAC minimally alters the result. However, when the baseline produces too few or too many instances, ISAC effectively corrects the output (see Figures 14a and 14b). These results show that ISAC reliably helps correct instance counts across diverse seeds for a fixed prompt, while leaving correct samples mostly unchanged.

These results highlight ISAC’s ability to consistently assist instance generation across diverse seeds for a fixed prompt—correcting incorrect instance counts while leaving already correct samples largely

Table 13: Additional quantitative results.

Method	Multi-Class Accuracy (\uparrow)					Multi-Instance Accuracy (\uparrow)					Latency (\downarrow)	VRAM (\downarrow)
	#2	#3	#4	#5	Average	#2	#3	#4	#5	Average		
SD1.4 (Rombach et al., 2022)	30%	2%	1%	0%	8%	94%	74%	28%	22%	55%	8s	4.9GB
+ A&E (Chefer et al., 2023)	50%	9%	8%	2%	17%	97%	79%	26%	23%	56%	17s	9.2GB
+ SynGen (Rassin et al., 2024)	54%	11%	6%	2%	18%	90%	69%	25%	19%	51%	19s	9.3GB
+ InitNO (Guo et al., 2024)	58%	10%	7%	4%	20%	94%	79%	31%	20%	56%	20s	9.6GB
+ TEBOpt (Chen et al., 2024a)	55%	13%	8%	2%	19%	91%	73%	31%	20%	54%	17s	9.3GB
+ ISAC (Ours)	66%	34%	29%	16%	36%	100%	90%	51%	40%	70%	21s	9.7GB
SD1.5 (Rombach et al., 2022)	28%	2%	1%	0%	8%	88%	65%	36%	26%	54%	12s	4.4GB
+ A&E (Chefer et al., 2023)	48%	10%	5%	2%	16%	91%	68%	34%	24%	54%	24s	9.1GB
+ SynGen (Rassin et al., 2024)	50%	9%	4%	2%	16%	84%	61%	38%	22%	51%	27s	9.2GB
+ InitNO (Guo et al., 2024)	55%	12%	7%	5%	20%	90%	68%	40%	29%	57%	29s	9.5GB
+ TEBOpt (Chen et al., 2024a)	52%	11%	8%	3%	18%	87%	65%	36%	27%	54%	25s	9.2GB
+ ISAC (Ours)	65%	31%	29%	18%	36%	95%	82%	56%	44%	69%	30s	9.6GB
SD2.1 (Rombach et al., 2022)	31%	6%	3%	0%	10%	91%	74%	41%	28%	58%	13s	4.8GB
+ A&E (Chefer et al., 2023)	53%	12%	4%	1%	18%	94%	79%	39%	29%	60%	26s	9.3GB
+ SynGen (Rassin et al., 2024)	55%	10%	7%	3%	19%	87%	69%	38%	25%	55%	29s	9.4GB
+ InitNO (Guo et al., 2024)	59%	13%	11%	5%	22%	91%	79%	44%	26%	60%	31s	9.7GB
+ TEBOpt (Chen et al., 2024a)	56%	14%	7%	6%	21%	88%	75%	44%	27%	58%	27s	9.4GB
+ ISAC (Ours)	67%	35%	34%	20%	39%	98%	88%	64%	42%	73%	32s	9.8GB
SDXL (Podell et al., 2023)	20%	4%	3%	0%	7%	90%	71%	49%	32%	61%	48s	12.8GB
+ ISAC (Ours)	57%	32%	29%	17%	34%	96%	89%	71%	47%	76%	101s	29.8GB
PixArt- α (Chen et al., 2023a)	27%	3%	1%	0%	8%	99%	93%	33%	15%	60%	17s	19.9GB
+ ISAC (Ours)	63%	30%	29%	21%	36%	100%	100%	56%	31%	72%	40s	53.7GB
PixArt- Σ (Chen et al., 2024b)	39%	8%	0%	0%	12%	98%	98%	30%	16%	60%	18s	19.9GB
+ ISAC (Ours)	78%	39%	31%	20%	42%	100%	100%	48%	31%	70%	41s	53.8GB
SD3.5-M (Esser et al., 2024)	62%	23%	12%	3%	25%	84%	71%	51%	51%	64%	40s	22.9GB
+ A&E (Chefer et al., 2023)	65%	29%	16%	5%	28%	86%	72%	52%	50%	65%	-	-
+ SynGen (Rassin et al., 2024)	66%	28%	15%	6%	28%	82%	68%	50%	48%	62%	-	-
+ InitNO (Guo et al., 2024)	77%	31%	17%	7%	33%	84%	73%	52%	49%	65%	-	-
+ TEBOpt (Chen et al., 2024a)	78%	31%	19%	8%	34%	85%	71%	52%	52%	65%	-	-
+ ISAC (Ours)	98%	51%	40%	20%	52%	98%	91%	72%	69%	83%	140s	74.8GB

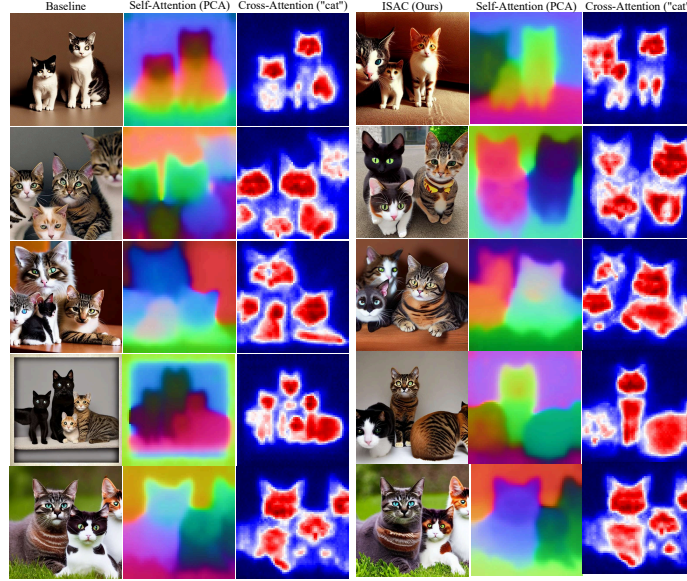
unchanged—demonstrating that the qualitative improvements align well with the gains observed in quantitative metrics across a wide range of classes beyond the animal examples discussed in the main text.

I FUTURE WORK

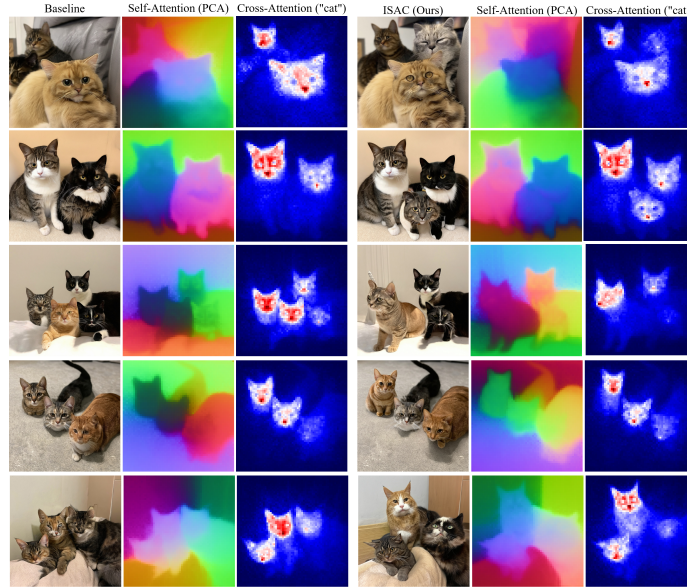
Extension to Distilled Diffusion Models. The timestep-wise instance-to-semantic dynamics observed in standard diffusion models appear to be compressed into a layer-wise instance-to-semantic dynamics in SDXL-Turbo (Sauer et al., 2024) (see Figure 15). This observation suggests that our method, ISAC, could potentially be extended to improve multi-class/multi-instance generation in distilled models (e.g., consistency models) as well.



Figure 13: Qualitative results beyond animals and simple prompts. SD1.5 (Rombach et al., 2022) with our method shows consistent success across variety of object classes. All prompts are taken from T2I-CompBench++ (Numeracy) (Huang et al., 2025) benchmark. 10 different seeds are used for each prompt to synthesize images.



(a) Result from Stable Diffusion v1.5 Rombach et al. (2022).



(b) Result from Stable Diffusion v3.5-M Esser et al. (2024).

Figure 14: Qualitative results across multiple seeds for the prompt *"three cats"*, which involves multiple instances of the same class. Our method (right) consistently generates images with the correct instance count, sharper object boundaries, and improved separation compared to the baseline (left).

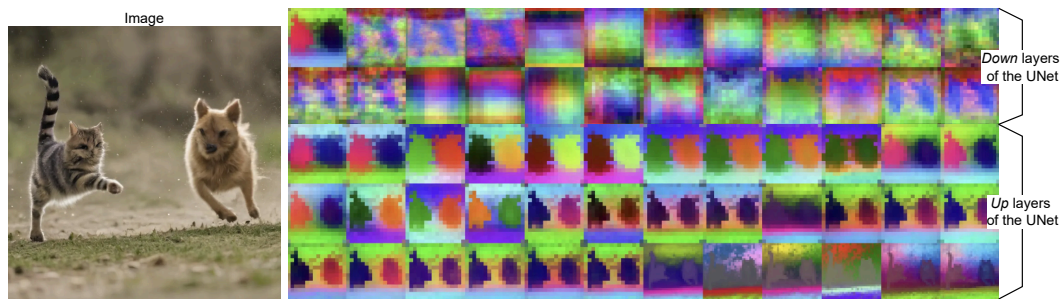


Figure 15: PCA visualizations of self-attention maps in the *down* and *up* layers of the UNet in SDXL-Turbo (Sauer et al., 2024). The timestep-wise instance-to-semantic dynamics observed in standard diffusion models appear to be compressed into a layer-wise instance-to-semantic dynamics in distilled models.