

Be Decisive: Noise-Induced Layouts for Multi-Subject Generation

OMER DAHARY, Tel Aviv University, Israel and Snap Research, Israel

YEHONATHAN COHEN, Tel Aviv University, Israel

OR PATASHNIK, Tel Aviv University, Israel and Snap Research, Israel

KFIR ABERMAN, Snap Research, United States of America

DANIEL COHEN-OR, Tel Aviv University, Israel and Snap Research, Israel



Fig. 1. Our method generates multi-subject images by leveraging the layout encoded in the initial noise. Having a layout control allows to accurately generate each subject. We predict the layout based on the initial noise, and refine it throughout the denoising process, aligning it with the prompt and making it more fine-grained. Using the layout encoded in the initial noise we preserve the prior of the original model and generate diverse compositions. Below each of the generated images, we show the layout predicted at three timesteps along the generation process.

Generating multiple distinct subjects remains a challenge for existing text-to-image diffusion models. Complex prompts often lead to subject leakage, causing inaccuracies in quantities, attributes, and visual features. Preventing leakage among subjects necessitates knowledge of each subject’s spatial location. Recent methods provide these spatial locations via an external layout control. However, enforcing such a prescribed layout often conflicts with the innate layout dictated by the sampled initial noise, leading to misalignment with the model’s prior. In this work, we introduce a new approach that predicts a spatial layout aligned with the prompt, derived from the initial noise, and refines it throughout the denoising process. By relying on this noise-induced layout, we avoid conflicts with externally imposed layouts and better preserve the model’s prior. Our method employs a small neural network to predict and refine the evolving noise-induced layout at each denoising step, ensuring clear boundaries between subjects while maintaining consistency. Experimental results show that this noise-aligned strategy achieves improved text-image alignment and more stable multi-subject generation compared to existing layout-guided techniques, while preserving the rich diversity of the model’s original distribution.

1 INTRODUCTION

Diffusion models have revolutionized the field of image synthesis, enabling the creation of high-quality and diverse images from intuitive conditions such as textual prompts. However, despite their significant success, these models still struggle to accurately align to complex prompts [Chefer et al. 2023]. Specifically, generating

multiple subjects remains surprisingly challenging, often resulting in inaccurate quantities, attributes, and visual features [Binyamin et al. 2024; Rassin et al. 2023; Yang et al. 2024].

Recent works have identified harmful leakage between subjects as a primary source to text-image misalignment. To address this issue, previous methods manipulate the denoising process by limiting inter-attention among distinct subjects [Dahary et al. 2025]. This approach requires knowing each subject’s spatial location, which is not explicitly represented within the model, and hence it relies on a prescribed layout control.

However, an externally imposed layout [Feng et al. 2024b,a; Qu et al. 2023; Yang et al. 2024; Zheng et al. 2023] can conflict with the layout implied by the sampled initial noise, creating tension with the model’s prior and potentially leading to inferior results or deviations from the model’s prior.

Specifically, as the image’s low frequencies are defined early in the denoising process, the initial noise plays a fundamental role in shaping the final layout of the generated image [Ban et al. 2024; Guo et al. 2024; Patashnik et al. 2023]. Therefore, steering the denoising trajectory toward a specific layout requires actively countering the model’s intrinsic prior, which naturally encodes a layout intent within the initial noise. This often pushes the generated image away from the image manifold, resulting in semantic misalignment and degradation of image quality.

“... science fiction movie poster with **two astronauts**,
a robot, and a **spaceship**”



Fig. 2. Our method generates images with multiple subjects without requiring external layout inputs. By following the innate noise-induced layout encoded in the sampled initial noise, we preserve the model’s prior and achieve diverse compositions. The second row show the initial noise-induced layout of the corresponding output images above. As can be seen, the initial layouts reflect the final composition of the generated images.

In this work, we introduce a method that derives a prompt-aligned spatial layout from the initial noise and iteratively refines it throughout the denoising process, as illustrated in Figure 1. By anchoring the layout around the initial noise, this approach stays consistent with the model’s prior, avoiding the conflicts introduced by externally imposed layouts. We argue that this approach promotes more natural and diverse compositions by minimizing resistance to the input noise, and hence succeeds in generating images that better adhere to the prompt.

To produce the layout, we train a small neural network that predicts the layout induced by the latent noisy image using features extracted from the denoising model. This network is applied throughout the denoising process, gradually refining the layout at each timestep to guide the generation toward layouts that remain both prompt-aligned and consistent across timesteps.

Our work embraces the motto “Be Decisive”. At each denoising step, we guide the process toward a well-defined layout, ensuring clear boundaries between subjects. In this approach, each subject is assigned to a distinct image region, preventing leakage and enhancing text-image alignment. Meanwhile, only minimal adjustments are made to the layout between steps, maintaining consistency with the noise-induced layout throughout the process.

Through extensive experiments, we demonstrate our method’s power in adhering to complex multi-subject prompts, and compare it with previous methods. Specifically, we demonstrate that our method generates combinations of classes, adjectives, and quantities while maintaining diverse layouts that are natural, as they remain consistent with the model’s prior layouts. Figure 2 highlights this diversity, showcasing compositions obtained by sampling different initial noises.

2 RELATED WORK

Diffusion models [Dhariwal and Nichol 2021; Podell et al. 2023; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022] have

achieved remarkable success in modeling the complex distribution of natural images. However, despite their advantages, these models still face limitations in adhering to detailed prompts, particularly those involving multiple subjects. Previous works have addressed challenges in multi-subject generation through two distinct approaches: conditioning the generation on a spatial layout or applying heuristics to attention maps to enforce the generation of each subject mentioned in the prompt.

Layout-Based Multi-Subject Generation. Layout-based methods have demonstrated greater consistency in multi-subject generation compared to text-to-image models. Early efforts incorporated layout information through techniques such as multiple diffusion compositions [Bar-Tal et al. 2023; Ge et al. 2023], guidance from model features [Kim et al. 2022; Luo et al. 2024; Voynov et al. 2023], specifically attention features [Chen et al. 2023a; Couairon et al. 2023; Kim et al. 2023; Liu et al. 2023; Phung et al. 2024; Xie et al. 2023], or fine-tuning [Avrahami et al. 2023; Li et al. 2023b; Nie et al. 2024; Yang et al. 2023; Zhang et al. 2023b].

Recent studies highlight the architectural tendency of attention layers to leak visual features between subjects – a phenomenon that complicates multi-subject generation [Dahary et al. 2025]. To address this, prior methods [Dahary et al. 2025; Wang et al. 2024a,b; Zhou et al. 2024] introduce techniques that mitigate such leakage by modifying the operation of attention layers within the model. However, these approaches rely on a predefined spatial layout to identify the subjects among which leakage should be prevented. In our work, we propose a method to dynamically define the spatial locations of subjects during image generation by extracting the layout throughout the process. This extracted layout is then used to prevent leakage, enabling the generation of accurate multi-subjects images.

To simplify the image generation process for users, a common practice is to automatically generate a layout prior to image generation. Several works leverage large language models (LLMs) for this task, employing in-context learning or chain-of-thought reasoning [Chen et al. 2023b; Feng et al. 2024b; Lian et al. 2023; Qu et al. 2023; Yang et al. 2024]. While these methods excel in producing plausible layouts, the separation between the prompt-to-layout and layout-to-image models often leads to inaccuracies or unnatural results in multi-subject images. Notably, Ranni [Feng et al. 2024a] proposes overcoming this limitation by jointly fine-tuning the LLM and diffusion model on a shared dataset. However, their approach demands significant resources, with results obtained using a large proprietary model trained on millions of examples.

Layout-Free Multi-Subject Generation. Numerous approaches have sought to address specific aspects of subject misalignment during inference without relying on a predefined spatial layout. Some manipulate text embeddings [Feng et al. 2022; Tunanyan et al. 2023], while others guide the model to disentangle the attention distributions of distinct subjects and attributes [Agarwal et al. 2023; Chefer et al. 2023; Li et al. 2023a; Meral et al. 2024; Rassin et al. 2023]. While these methods show some success, their effectiveness often hinges on the initial noise, resulting in unstable outcomes. To enhance robustness, other techniques [Bao et al. 2024; Wang et al. 2023] employ fine-tuning based on similar heuristics in a self-supervised

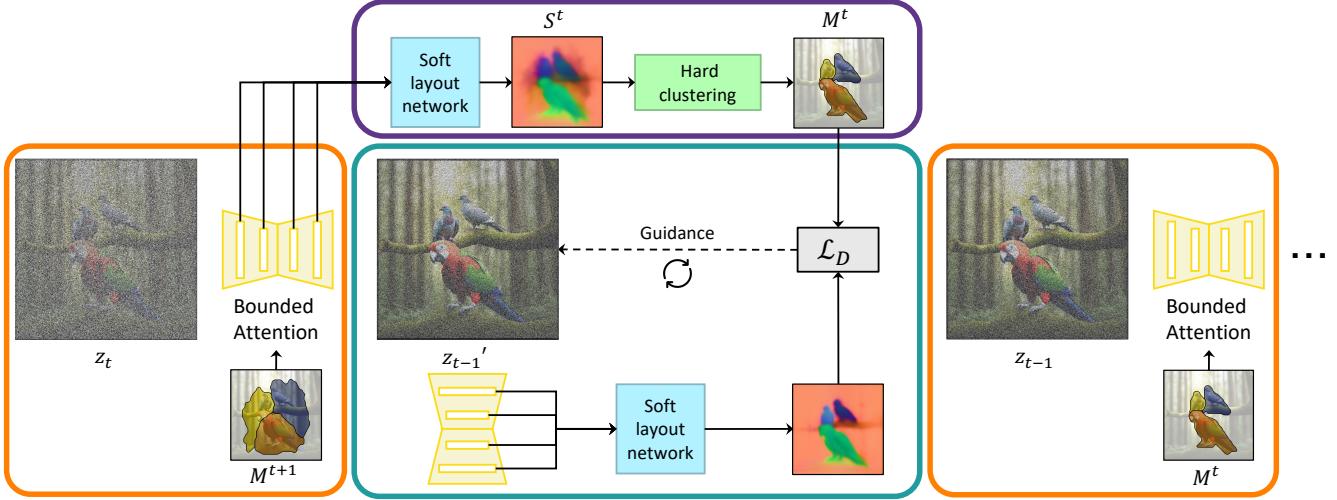


Fig. 3. Our method steers the denoising process by applying iterative guidance (turquoise box) after each denoising step (orange regions). At denoising step t (left orange box), we predict a soft-layout S^t based on the diffusion model’s features, and cluster it to form a hard-layout M^t (purple box). This hard-layout is then used to control the layout of the next denoising step (right orange box). In the guidance stage, we optimize the latent image, with the objective to align its associated updated soft-layout with the hard-layout M^t .

manner. Nevertheless, due to the model’s limitations in interpreting quantities and distinguishing between numerous subjects, these approaches often struggle to generate more than two or three distinct subjects and fail to support multiple instances of the same class effectively.

Other works specifically tailor solutions for accurate subject quantities [Binyamin et al. 2024; Kang et al. 2023; Zhang et al. 2023a]. While these methods perform well for single-class scenarios, they lack the generality needed for complex compositions involving multi-class subjects and attributes. In contrast, our approach provides comprehensive control over multi-class subjects, quantities, and attributes, addressing the limitations of existing layout-free methods.

3 PRELIMINARY: BOUNDED ATTENTION

Text-to-image diffusion models struggle to generate accurate multi-subject images due to visual leakage between subjects. Prior work [Dahary et al. 2025] identified the model’s attention layers as the primary source of this leakage — where features of semantically similar subjects are indiscriminately blended — and proposed Bounded Attention as a training-free solution to mitigate it.

Given an input layout, Bounded Attention modifies the attention layers during the denoising process by masking the attention between queries and keys of different subjects. In cross-attention layers, it constrains each subject’s attention to its corresponding textual tokens. In self-attention layers, it restricts attention to pixels within the subject’s own region and the background, explicitly excluding other subjects. This masking scheme reduces the influence of irrelevant visual and textual tokens on each pixel, maintaining the distinct visual features of each subject.

During generation, Bounded Attention alternates between denoising steps and guidance steps, both of which adopt the masking

scheme. In guidance mode, the latent representation is optimized to adhere to the input layout: $z_t^{\text{opt}} = z_t - \beta \nabla_{z_t} (\mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{self}})$, where $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{self}}$ are loss terms that encourage the respective cross- and self-attention maps to focus within each subject’s designated mask. By isolating attention for each subject, the masking scheme avoids guidance artifacts caused by forcing similar queries to diverge, maintaining a trajectory that is better aligned with the data manifold.

In our work, we adopt Bounded Attention’s masking scheme to reduce leakage, but instead of relying on a prescribed layout, we extract the noise-induced layout and refine it between denoising steps. We further modify the guidance procedure to promote decisiveness — that is, enforcing strict subject boundaries throughout the layout refinement process.

4 METHOD

Our method aims to facilitate the generation of multiple distinct subjects using an existing text-to-image model [Podell et al. 2023]. We steer the denoising process to adhere to a layout that allows preventing unwanted leakage among the subjects. Our key idea is to progressively define a prompt-aligned spatial layout based on features extracted from the noisy latent images along the denoising process. We then encourage the denoising process to follow these layouts, upholding this initial “decision”.

Figure 3 illustrates the overall structure of our inference pipeline. Our method is built on a denoising process to which we apply Bounded Attention [Dahary et al. 2025] (marked in orange boxes) controlled by layout masks M^t . We add two components to the denoising process. First, a component that predicts a prompt-aligned layout M^t from a noisy latent image z_t based on features extracted from the diffusion model (purple box). Second, a guidance mechanism that optimizes a noisy latent image so that its induced layout

aligns with the previous layout (turquoise box). This mechanism encourages a “decisive” generation process, where each subject mentioned in the prompt is consistently assigned to its own distinct image region across timesteps.

Both components rely on a *soft-layout* S^t . The soft-layout is a timestep-dependent feature map that reflects the likelihood that two pixels will be associated with a common subject. In the following, we elaborate on the soft-layout and its use.

4.1 Soft-Layout

We begin by explaining the motivation behind our soft-layouts. Extracting fine-grained layouts directly from the initial noise is inherently challenging since the image is formed in a gradual manner. Moreover, predicted layouts might not perfectly correspond to the subjects specified in the prompt. To address these challenges, we introduce the notion of soft-layout, a feature map that represents each pixel as a descriptor encapsulating its potential to associate with other pixels in composing a single subject. In the first timesteps, due to high uncertainty, the soft-layout encodes a coarse layout. At later timesteps, the soft-layout is more granular and precise. Our use of the soft-layout is two-fold. First, it is used to predict the masks M^t , termed as *hard-layout*, which bounds the attention in the denoising steps. Second, we optimize the noisy latent image to produce a soft-layout that agrees with M^t .

At the top of Figure 4, we display the progressive layouts produced by our full pipeline. In the middle, we show the corresponding layouts without guidance. As illustrated, guidance is crucial for maintaining consistent hard-layouts across timesteps, thereby facilitating convergence to a prompt-aligned layout by the end of the denoising process.

We now turn to formally define the soft-layout and elaborate on the network we train to predict it. A soft-layout $S^t \in \mathbb{R}^{n \times d}$ is a feature map, encoding n pixels as d -dimensional vectors, where the similarity of two feature vectors $S^t[x_1], S^t[x_2]$ indicates correspondence to the same subject in the generated image. To produce the soft-layout, we train a network that takes as input a set of features extracted from various layers of the diffusion model.

Dataset. To train our network, we automatically construct a small dataset of ~ 1500 images synthesized by the diffusion model, along with their segmentation maps. First, we randomly generate a set of prompts specifying multiple subject classes and their quantities (see full details in the supplemental). Then, we synthesize images based on these prompts, and segment them by feeding the corresponding subject names to GroundedSAM [Ren et al. 2024b]. We filter out ambiguous examples, where two segmentation masks share a large overlap, and select a single label for each segment based on the segmentation model’s confidence score.

Notably, we do not apply any filtering based on prompt alignment. This allows the network to predict soft-layouts that match the diffusion model’s intent, even if it does not adhere to the prompt. In turn, this enables our guidance mechanism to detect misalignments early in the denoising process and apply corrective updates to the latent.

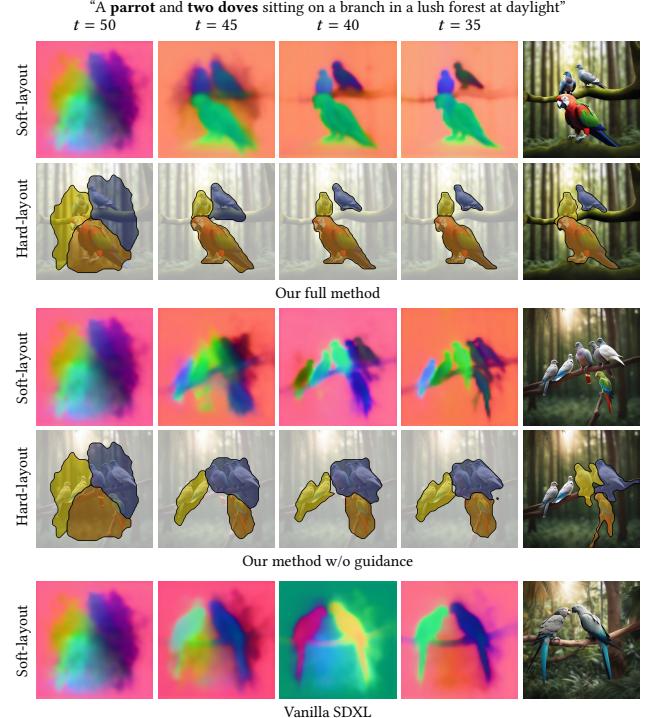


Fig. 4. The figure illustrates the progression of the soft- and hard-layouts in three cases. The top row shows results from our full method. The middle row presents our method without guidance. The bottom row shows vanilla SDXL, where only the soft-layout extracted from the noisy latents is displayed. Below each image, we show the hard-layout obtained at the final timestep.

Architecture. Following Readout Guidance [Luo et al. 2024], we design our model as a collection of lightweight convolutional heads, each processing different features from the denoising model along with the current time embedding. The outputs of these heads are then averaged using learnable weights, and fed into a convolutional bottleneck head, which outputs a $64 \times 64 \times 10$ feature map, representing the soft-layout.

We attach our heads to the attention layers, which are known to be highly indicative of the image structure and subject boundaries [Hertz et al. 2022; Patashnik et al. 2023; Tumanyan et al. 2023]. Specifically, we use the cross-attention queries and the self-attention keys at the decoder layers. See the supplemental for full architectural details.

Training. We train the soft-layout network with a triplet loss [Schroff et al. 2015], encouraging feature similarity between pixels of the same segment, and dissimilarity between different segments.

Formally, given a random timestep t and an image with k subject segments $\{M_j\}_{j=1}^k$ and a background M_0 , we sample triplets of pixel coordinates $x_{i_a}, x_{i_p} \in M_{j_p}$ and $x_{i_n} \in M_{j_n}$, where $M_{j_p} \neq M_{j_n}$. Then, we compute the following loss

$$\sum_i \left[\text{sim}(S^t[x_{i_a}], S^t[x_{i_n}]) - \text{sim}(S^t[x_{i_a}], S^t[x_{i_p}]) + \alpha \right]_+, \quad (1)$$

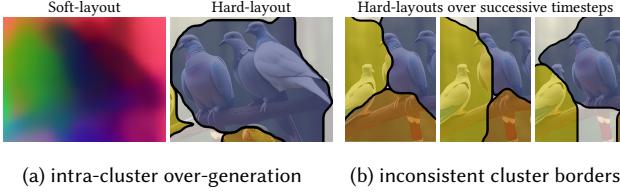


Fig. 5. Without guidance, we observe two types of layout failures: (a) intra-cluster over-generation, where multiple subjects are assigned to a single cluster due to high variance in the soft-layout; and (b) inconsistent cluster borders across timesteps, leading to subject over-generation and leakage caused by oscillating boundaries.

where sim is the cosine-similarity, α is the similarity margin between positive and negative samples, and $[\cdot]_+$ is the ReLU operation.

4.2 From Soft to Hard Layouts

While the soft-layout represents the original model’s future intent, to successfully generate multiple prompt-aligned subjects, it is necessary to uphold clear subject boundaries in accordance to the prompt. To achieve this, we derive a hard-layout from the soft-layout produced by our network.

More specifically, given k subjects mentioned in the prompt, we apply K-Means to cluster the soft-layout into $k + 1$ segments: k for the subjects, and one for the background. We set the background M_0 as the cluster that has the biggest overlap with the image’s border, and recursively cluster each of the other segments into two sub-clusters, continuing the process with the bigger sub-cluster, until the variance is smaller than $\sigma_{\text{cluster}}^2$. Any sub-cluster dropped during this process is added to M_0 .

Finally, we must tag each subject cluster with an appropriate label representing a specific subject instance. After the first denoising step, at $t = T$, we compute the average cross-attention map of each subject noun [Epstein et al. 2023; Hertz et al. 2022; Patashnik et al. 2023] and use the Hungarian algorithm to assign instances to clusters such that the corresponding cross-attention response in each cluster is maximized.

To avoid leakage, our initial decision regarding each subject’s location must be respected throughout the generation process. Thus, for $t < T$, we stack the soft-layout S^t with the previous soft-layouts $S^{t+1}, \dots, S^{\min(t+w, T)}$ from w earlier timesteps, before performing hard-clustering. Since clusters may shift over time, we reassign their labels at each timestep using the Hungarian algorithm, matching each cluster in M^t to a cluster in M^{t+1} such that their intersection-over-union (IoU) is maximized.

4.3 Decisive Guidance

To encourage decisiveness – in the sense of maintaining consistent subject boundaries throughout generation – we perform guidance steps after each denoising step. These steps optimize the intermediate latent z_{t-1} to align the predicted soft-layout S^{t-1} with the previous hard-layout M^t (see turquoise box in Figure 3).

First, to integrate each subject’s semantics to its designated segment in M^t , we apply the cross-attention loss $\mathcal{L}_{\text{cross}}$ from Bounded

Attention (see Section 3). Notably, Bounded Attention also introduces a self-attention loss $\mathcal{L}_{\text{self}}$ to mitigate subject neglect in fixed, externally provided layouts by discouraging background attention. However, we omit this term, as it is unnecessary with our noise-induced layouts and its removal significantly reduces runtime.

Nonetheless, our evolving layouts give rise to two distinct failure modes. These are illustrated in Figure 5, which presents zoomed-in views of the soft- and hard-layouts generated without guidance (originally shown in Figure 4). On the left (Figure 5a), the soft-layout contains three spatially separated foreground regions (colored green, purple, and dark red) within a single hard-cluster, each corresponding to a distinct dove. On the right (Figure 5b), the middle dove is generated at the intersection of three hard-clusters and does not maintain consistent membership in any single cluster during denoising. As a result, its lower body is initially assigned to the parrot cluster, leading to a hybrid generation in which the dove inherits a parrot-like tail.

To address the first issue (Figure 5a), we introduce a variance loss \mathcal{L}_{var} that encourages low cluster variance in S^{t-1} with respect to the previous hard-layout M^t :

$$\mathcal{L}_{\text{var}} = \frac{1}{k+1} \sum_{j=0}^k \frac{1}{|M_j^t|} \sum_{x_i \in M_j^t} \text{sim}^2(S^{t-1}[x_i], \mu_j^{t-1}), \quad (2)$$

where μ_j^{t-1} is the mean soft-layout feature vector of cluster j :

$$\mu_j^{t-1} = \frac{1}{|M_j^t|} \sum_{x_i \in M_j^t} S^{t-1}[x_i]. \quad (3)$$

This loss promotes intra-cluster similarity, encouraging each cluster to represent a coherent subject instance.

To avoid cluster boundaries from oscillating between timesteps (Figure 5b), we compute the Dice segmentation loss $\mathcal{L}_{\text{dice}}$ [Millettari et al. 2016] between the hard-layout M^t and a probabilistic layout $P^{t-1} \in \mathbb{R}^{n \times (k+1)}$, where each element $P^{t-1}[x_i, j]$ represents the probability that pixel x_i belongs to cluster j :

$$P^{t-1}[x_i, \cdot] = \text{softmax}\left(\left\{\text{sim}(S^{t-1}[x_i], \mu_j^{t-1}) / \tau\right\}_{j=0}^k\right) \in \mathbb{R}^{k+1}, \quad (4)$$

where τ is a temperature hyperparameter. This term penalizes ambiguous pixel-cluster associations, promoting sharper and more consistent cluster boundaries.

Together, these three terms address complementary aspects of the layout refinement process: $\mathcal{L}_{\text{cross}}$ promotes the proper semantic alignment in each cluster, \mathcal{L}_{var} reduces intra-cluster ambiguity, and $\mathcal{L}_{\text{dice}}$ encourages temporal consistency and boundary sharpness. The final decisiveness loss is defined as:

$$\mathcal{L}_{\text{decisive}} = \alpha_{\text{cross}} \mathcal{L}_{\text{cross}} + \alpha_{\text{var}} \mathcal{L}_{\text{var}} + \alpha_{\text{dice}} \mathcal{L}_{\text{dice}}, \quad (5)$$

where $\alpha_{\text{cross}}, \alpha_{\text{var}}, \alpha_{\text{dice}}$ are the respective weighting coefficients.

Ablation studies evaluating the contribution of each component are provided in the supplementary material.

5 EXPERIMENTS

In this section, we present both qualitative and quantitative experiments to evaluate the effectiveness of our method. We compare



Fig. 6. Generated images across different seeds. Our method follows the noise-induce layouts to generate prompt-aligned images with diverse compositions.

our approach against four training-free baseline methods: Make-It-Count (MIC) [Binyamin et al. 2024], RPG [Yang et al. 2024], Attend-and-Excite (A&E) [Chefer et al. 2023], and Bounded Attention (BA) [Dahary et al. 2025]. Since BA operates on layouts, we use an LLM to automatically provide it with layouts constructed from given prompts (denoted as LLM+BA). Furthermore, we include comparisons with LMD+[Lian et al. 2023] and Ranni [Feng et al. 2024a], which require training.

5.1 Qualitative Results

Layout diversity. We begin our experiments by showing the effectiveness of our method in generating diverse and natural layouts that adhere to the prompt. Each row of figures 6,10 depict images generated from a single prompt using different random seeds. As can be seen, our results exactly match subject descriptions, displaying proper combinations of classes, attributes and quantities, while still demonstrating unique and believable compositions.

Non-curated results. We conduct a non-curated comparison with our baseline in Figure 11 by sampling each method seven times, using a single prompt and the seeds 0 to 6. We also display the results obtained by Flux.

While LLM+BA is able to generate correct images four out of seven times, our method is able to correctly adhere to the prompt in each image without requiring an input layout. Notably, none of the other methods, including Flux, are able to generate even one sample that match the prompt, often depicting subject amalgamations due to severe leakage. Specifically, SDXL, LLM+BA and A&E suffer from over-generation of subjects, while Flux, RPG and Ranni struggle due to under-generation. On the other hand, LMD+ is able to construct the correct quantities, but is prone to generating unnatural compositions, where subjects appear disjointed from the background.

Multiple Personalized Subjects. Leakage between subjects is particularly noticeable when generating personalized subjects. Here, we show that our method can be seamlessly integrated with an existing



Fig. 7. Results of integrating our method with an existing personalization method, enabling the generation of multiple personalized individuals within the same image.

personalization method to facilitate the generation of multiple personalized subjects. Specifically, we utilize a method trained to generate specific individuals by injecting personalized features through the cross-attention layers of a text-to-image model [Patashnik et al. 2025]. This method does not inherently support the generation of two individuals, as demonstrated in Figure 7. However, combining this method with ours, enables the accurate generation of diverse images with multiple individuals.

Comparisons with baselines. We present a qualitative comparison in Figure 8. All other methods struggle to generate multi-subjects prompts due to leakage. In the first row, none of the competing methods are able to generate the distinct characteristics of each of the bears. In the second row, they either generate the wrong number of subjects, or leak the colors of the carpet or the cars into the teddy bears. Specifically, the current LLM-based methods exhibit either subpar control over subject quantities (LLM+BA, RPG, Ranni), or unnatural grid-like subject arrangements (LLM+BA, LMD+). In comparison, our method successfully generates prompt-aligned images with natural-looking compositions.

5.2 Quantitative Results

Dataset evaluation. We perform quantitative evaluation on the T2I-CompBench dataset [Huang et al. 2023], assessing our method’s performance across the following key aspects: multi-class compositions, attribute binding, and numeracy. We further measure layout diversity, which quantifies the variability of generated compositions across different seeds. We summarize the results in Figure 9, and refer to the supplement for the full table.



Fig. 8. **Qualitative comparison** of our method with baseline methods. We provide more examples in the supplement.

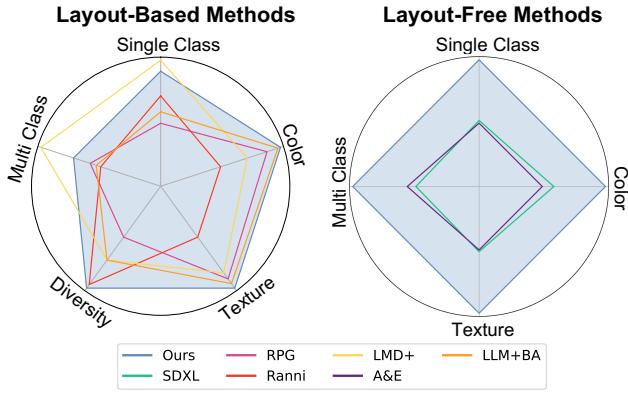


Fig. 9. Quantitative comparison of our method against layout-based and layout-free methods. Results demonstrate that while other methods face trade-offs between metrics, our method consistently achieves high scores across all metrics.

To measure layout diversity, we sample 20 random prompts from CompBench’s single-class dataset and generate five images per prompt using five random seeds. We align all five layouts by maximizing the IoU between them using the Hungarian algorithm. Diversity is quantified as the average $1 - \text{IoU}$ between all layout pairs of the same prompt. As reported, our method achieves significantly higher diversity than the baseline, preserving the innate variability of the model’s prior, in contrast to the limited diversity of LLM-based methods.

All other metrics were assessed on 200 prompts, sampled from the respective category in CompBench. Color and texture binding was evaluated using BLIP-VQA [Huang et al. 2023], while single-class and multi-class compositions were evaluated using the F1 scores between ground-truth subject quantities and the quantities computed by GroundedSAM [Ren et al. 2024a] on the generated images.

Table 1. User study results.

	SDXL	LLM+BA	RPG	Ranni	LMD+	MIC
Our score vs.	0.74	0.87	0.96	0.89	0.9	0.53

While other methods are tailored towards enhancing specific metrics, our approach consistently achieves high performance across all measurements, surpassing competitors in most cases.

Lastly, since MIC is limited to single-class prompts, we only measure its performance on this specific metric. Our method achieves a score of 0.837, compared to MIC’s score of 0.772.

User study. The automatic metrics in Figure 9 fail to detect semantic leakage, as they rely on models trained on real images, where such issues do not arise. To address this limitation, we conduct a user study. We utilize ChatGPT to generate 25 prompts enlisting three to four visually similar, but distinct, animals, with an appropriate background. For each prompt, participants were shown 10 images: five generated by our method, and five generated by a competing method. Users were then tasked with selecting images with realistic compositions that accurately reflect the prompt. For evaluation against MIC, we additionally generate five prompts with single-class quantities. We collected 192 responses from 32 participants. Table 1 reports the conditional probability of a selected image being generated by our method versus competitors. The results showcase our method’s superiority in handling complex multi-subject prompts, with our scores substantially improving over layout-based methods. Notably, even though MIC is specifically designed to tackle single-class quantities, our method receives comparable scores, while still being versatile enough to support more complex prompts.

6 CONCLUSIONS

We have addressed the notorious difficulty of generating multiple distinct subjects from complex prompts in text-to-image diffusion models. Recognizing that inter-subject leakage is the primary issue,

and that bounding mutual attention offers a viable solution, we designed a mechanism to define a layout for controlling inter-subject attention. Our key contribution lies in using the natural latent layout defined by the initial noise of the model, rather than imposing an external layout. By making only small adjustments on-the-fly, our approach remains rooted in the original distribution of the model, benefiting from denoising a signal already close to that distribution. Empirical evaluations confirm that this strategy provides a stronger balance between text-image alignment and visual diversity compared to layout-driven alternatives.

It is important to recognize that the multi-subject generation problem is intrinsically tied to the pretrained model's prior. When the underlying network has not been sufficiently exposed to images featuring multiple distinct subjects, its learned distribution may be ill-equipped to handle complex multi-subject arrangements. As a result, any approach aiming to improve multi-subject generation, ours included, must contend with these fundamental distributional constraints. Although our method outperforms existing alternatives, there remains a ceiling imposed by the model training data, restricting how effectively multi-subject prompts can be addressed in practice.

The main limitation of our method lies in the computational cost of the iterative guidance and its tendency to push the optimized latent away from the prior distribution. In the future, we aim to explore regularization techniques to keep the latent closer to its original distribution or replace the optimization process with feature injection from a control map representing the target clusters.

ACKNOWLEDGMENTS

We thank Elad Richardson and Narek Tumanyan for their early feedback and helpful suggestions. We also thank the anonymous reviewers for their meticulous comments which have helped improve our work. This work was partially supported by ISF (grant 3441/21).

REFERENCES

- Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2023. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2283–2293.
- Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18370–18380.
- Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Boqing Gong, Cho-Jui Hsieh, and Minhao Cheng. 2024. The Crystal Ball Hypothesis in diffusion models: Anticipating object positions from initial noise. *arXiv preprint arXiv:2406.01970* (2024).
- Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. 2024. Separate-and-Enhance: Compositional Finetuning for Text-to-Image Diffusion Models. In *ACM SIGGRAPH 2024 Conference Papers*. 1–10.
- Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation. (2023).
- Lital Binyamin, Yoav Tewel, Hilit Segev, Eran Hirsch, Royi Rassim, and Gal Chechik. 2024. Make It Count: Text-to-Image Generation with an Accurate Number of Objects. *arXiv preprint arXiv:2406.10210* (2024).
- Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Golokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, et al. 2024. Getting it right: Improving spatial consistency in text-to-image models. In *European Conference on Computer Vision*. Springer, 204–222.
- Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.
- Minghao Chen, Iro Laina, and Andrea Vedaldi. 2023a. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373* (2023).
- Xiaohui Chen, Yongfei Liu, Yingxiang Yang, Jianbo Yuan, Quanzeng You, Li-Ping Liu, and Hongxia Yang. 2023b. Reason out your layout: Evoking the layout master from large language models for text-to-image synthesis. *arXiv preprint arXiv:2311.17126* (2023).
- Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2174–2183.
- Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. 2025. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*. Springer, 432–448.
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. 2023. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems* 36 (2023), 16222–16239.
- Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032* (2022).
- Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024b. Layoutpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems* 36 (2024).
- Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. 2024a. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4744–4753.
- Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. 2023. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7545–7556.
- Xiefan Guo, Jinlin Liu, Miaoamiao Cui, Jinkai Li, Hongyu Yang, and Di Huang. 2024. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9380–9389.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).
- Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-combench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023), 78723–78747.
- Wonjun Kang, Kevin Galim, and Hyung Il Koo. 2023. Counting guidance for high fidelity text-to-image synthesis. *arXiv preprint arXiv:2306.17567* (2023).
- Gyeonghyeon Kim, Wooseok Jang, Gyuseong Lee, Susung Hong, Junyoung Seo, and Seungryong Kim. 2022. Dag: Depth-aware guidance with denoising diffusion probabilistic models. *arXiv preprint arXiv:2212.08861* (2022).
- Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. 2023. Dense Text-to-Image Generation with Attention Modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7701–7711.
- Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. 2023a. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864* (2023).
- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.
- Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655* (2023).
- Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023. Customizable image synthesis with multiple subjects. *Advances in neural information processing systems* 36 (2023), 57500–57519.
- Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. 2024. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8217–8227.
- Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. 2024. 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*. 9005–9014.
- Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 565–571.
- Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. 2024. Compositional Text-to-Image Generation with Dense Blob Representations. *arXiv preprint arXiv:2405.08246* (2024).



Fig. 10. Generated images across different seeds. Our method follows noise-induce layouts to generate diverse compositions, while still faithfully depicting subject characteristics such as class features, attributes and quantities. Note the rich layout diversity of the results.

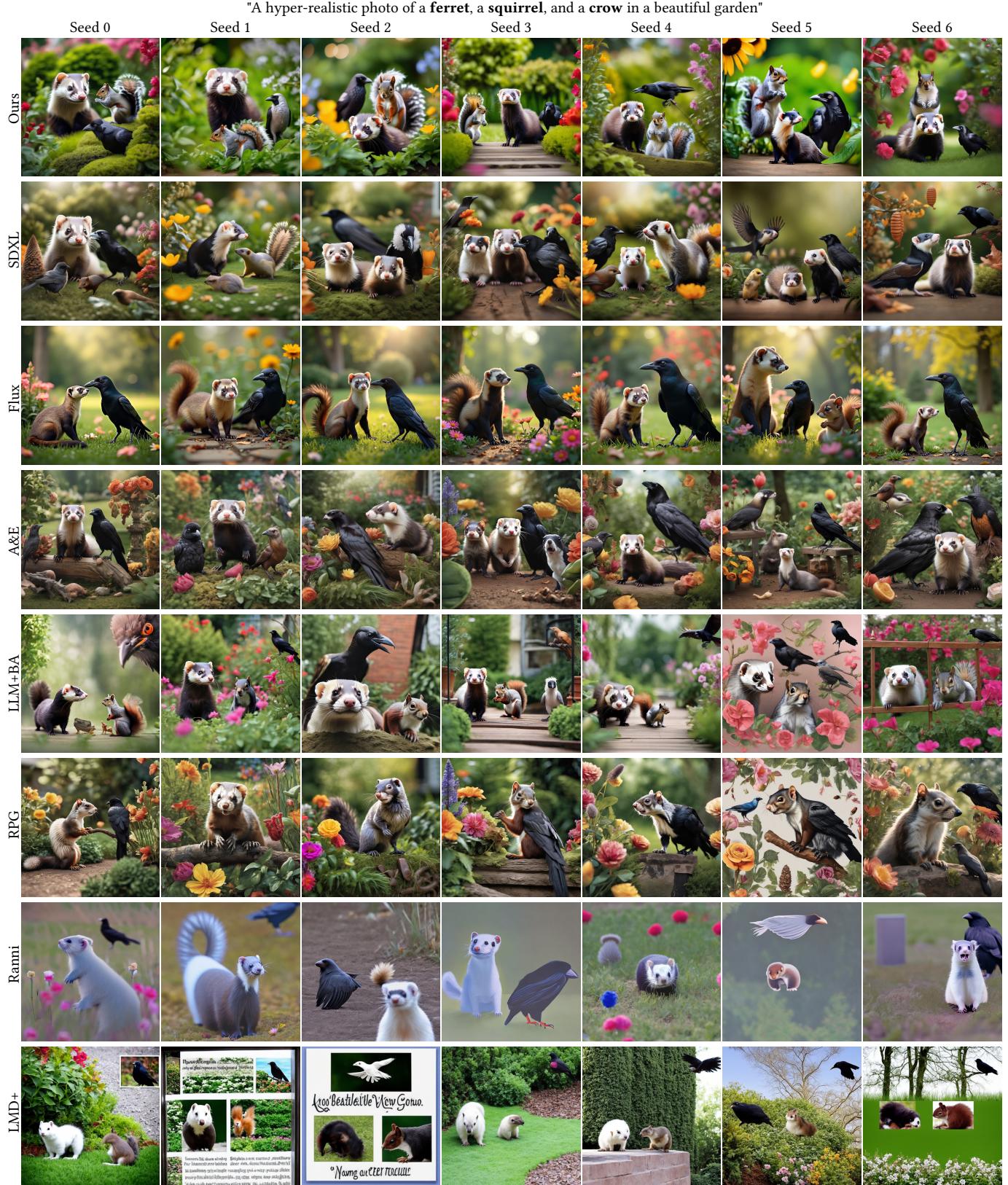


Fig. 11. Comparison of non-curated images generated from seeds 0 to 6.

- Or Patashnik, Rinon Gal, Daniil Ostanhev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 2025. Nested Attention: Semantic-aware Attention Values for Concept Personalization. arXiv:2501.01407 [cs.CV]
- Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23051–23061.
- Quynh Phung, Songwei Ge, and Jia-Bin Huang. 2024. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7932–7942.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
- Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. LayoutLLM-T2I: Eliciting Layout Guidance from LLM for Text-to-Image Generation. *arXiv preprint arXiv:2308.05095* (2023).
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022). 3.
- Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment. *arXiv preprint arXiv:2306.08877* (2023).
- Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024a. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).
- Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, ZhaoYang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024b. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *ArXiv abs/2401.14159* (2024). <https://api.semanticscholar.org/CorpusID:267212047>
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
- Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.
- Hazarapet Tunanyan, Dejia Xu, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. 2023. Multi-Concept T2I-Zero: Tweaking Only The Text Embeddings and Nothing Else. *arXiv preprint arXiv:2310.07419* (2023).
- Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
- Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. 2024a. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6232–6242.
- Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. 2024b. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209* (2024).
- Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. 2023. Tokencompose: Grounding diffusion with token-level supervision. *arXiv preprint arXiv:2312.03626* (2023).
- Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. 2023. BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7452–7461.
- Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*.
- Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linji Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14246–14255.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
- Ruisu Zhang, Yicong Chen, and Kangwook Lee. 2023a. Zero-shot Improvement of Object Counting with CLIP. In *Ro-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.
- Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22490–22499.
- Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. 2024. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6818–6828.

APPENDIX

A TECHNICAL DETAILS

A.1 Architecture and Hyperparameters

In our experiments, we used SDXL [Podell et al. 2023] as the backbone model.

Our soft-layout network follows the Readout Guidance architecture with spatially aligned heads [Luo et al. 2024], with minor modifications:

- We use the self-attention keys and cross-attention queries from the decoder layers as inputs to the network, adjusting the input channels accordingly.
- We modify the final convolutional layer to output 10 channels instead of 3.

The network was trained for 5,000 steps using a learning rate of 10^{-4} . To compute the triplet loss, we sample 50 pixel triplets per image, selecting subject pixels with a probability of 0.75 and background pixels with a probability of 0.25. A similarity margin of $\alpha = 0.5$ is applied.

During image generation, we use 50 denoising steps, a DDPM scheduler, and a guidance scale of 7.5. Guidance is applied during the first 15 denoising steps, with 5 gradient descent iterations per step. The decisive loss is computed using the weights $\alpha_{cross} = 0.3$, $\alpha_{var} = 0.21$, and $\alpha_{dice} = 0.49$. For \mathcal{L}_{dice} , a temperature value of $\tau = 15$ is used. For hard-clustering, we set a sliding window size of $w = 30$ and a variance threshold of $\sigma_{cluster}^2 = 0.025$.

A.2 Dataset Generation

Our dataset consists of approximately 1,500 generated images and their corresponding segmentation maps.

To construct training prompts, we use the same 20 MSCOCO classes as in MIC [Binyamin et al. 2024]. Each prompt randomly includes 1–3 classes. For each selected class, we assign a quantity between 1 and 10, with a probability of 0.9. We optionally prepend a prefix (with probability 0.8) and append a postfix (with probability 0.6), both sampled from fixed lists:

- Prefixes: “a photo of”, “an image of”, “a picture of”, “a painting of”.
- Postfixes: “on the grass”, “on the road”, “on the ground”, “in a yard”.

We observe that our soft-layout network generalizes well to unseen object classes, backgrounds, and prompt structures, owing to its lightweight design and the use of expressive attention features from the pre-trained diffusion model.

A.3 Computational Resource Usage

All experiments were conducted on an NVIDIA A100 GPU, with all computations – including clustering – performed on the GPU. Similar to Readout Guidance, our sampling process takes approximately 77 seconds and utilizes 36 GB of VRAM, compared to 7 seconds and 8 GB for vanilla SDXL. Our implementation builds upon the Readout Guidance and Bounded Attention codebases, which were not optimized for resource efficiency. As such, further code optimization is likely to reduce both runtime and memory usage.

Table 2. Quantitative evaluation.

Method	Color	Texture	Single-Class	Multi-Class	Layout	Diversity
Ours	0.704	0.686	0.837	0.723	0.718	
SDXL	0.568	0.660	0.746	0.676	-	
A&E	0.537	0.659	0.742	0.682	-	
LLM+BA	0.685	0.665	0.659	0.603	0.408	
RPG	0.604	0.643	0.609	0.635	0.155	
Ranni	0.259	0.445	0.729	0.579	0.679	
LMD+	0.457	0.614	0.885	0.898	0.408	

Table 3. Ablation user study results.

Method	Prompt-Alignment Accuracy
w/o $\mathcal{L}_{decisive}$	0.016
w/o \mathcal{L}_{cross}	0.442
w/o \mathcal{L}_{var}	0.447
w/o \mathcal{L}_{dice}	0.105
$\mathcal{L}_{decisive}(S^{t-1}, M^{t-1})$	0.289
Full method	0.832

B ADDITIONAL RESULTS

We use the same baseline as in the main paper: LLM+BA [Dahary et al. 2025], RPG [Yang et al. 2024], Ranni [Feng et al. 2024a], LMD+[Lian et al. 2023], and A&E [Chefer et al. 2023].

B.1 Quantitative Results

Table 2 presents the quantitative results comparing our method with the baseline. Unlike other approaches that balance trade-offs between metrics, our method consistently delivers high performance across all metrics.

B.2 Qualitative Results

Figure 12 showcases additional qualitative comparison results. Unlike competing methods, which fail to accurately generate all subjects from the prompt, our method consistently preserves the intended semantics of each subject. For instance, in the first row, none of the methods successfully generate all the fruits specified in the prompt. Similarly, in the last row, none of the methods accurately capture all the animals in the prompt, with most suffering from attribute leakage.

B.3 Ablation Studies

Quantitative Evaluation. To quantitatively assess the contribution of each component, we conducted a user study, following the same format as our benchmark user study, and composed of a subset of 10 random prompts from the full user study. We report the percentage of user-selected images for each methods, i.e. prompt-alignment accuracy, as recorded by 19 participants, in Table 3.

Qualitative Evaluation. We display qualitative ablation studies in Figure 13, where we systematically vary our method’s configuration to assess each component’s importance.

As can be seen in the leftmost four columns, neglecting our decisive guidance, or any of its terms, promotes subject over-generation

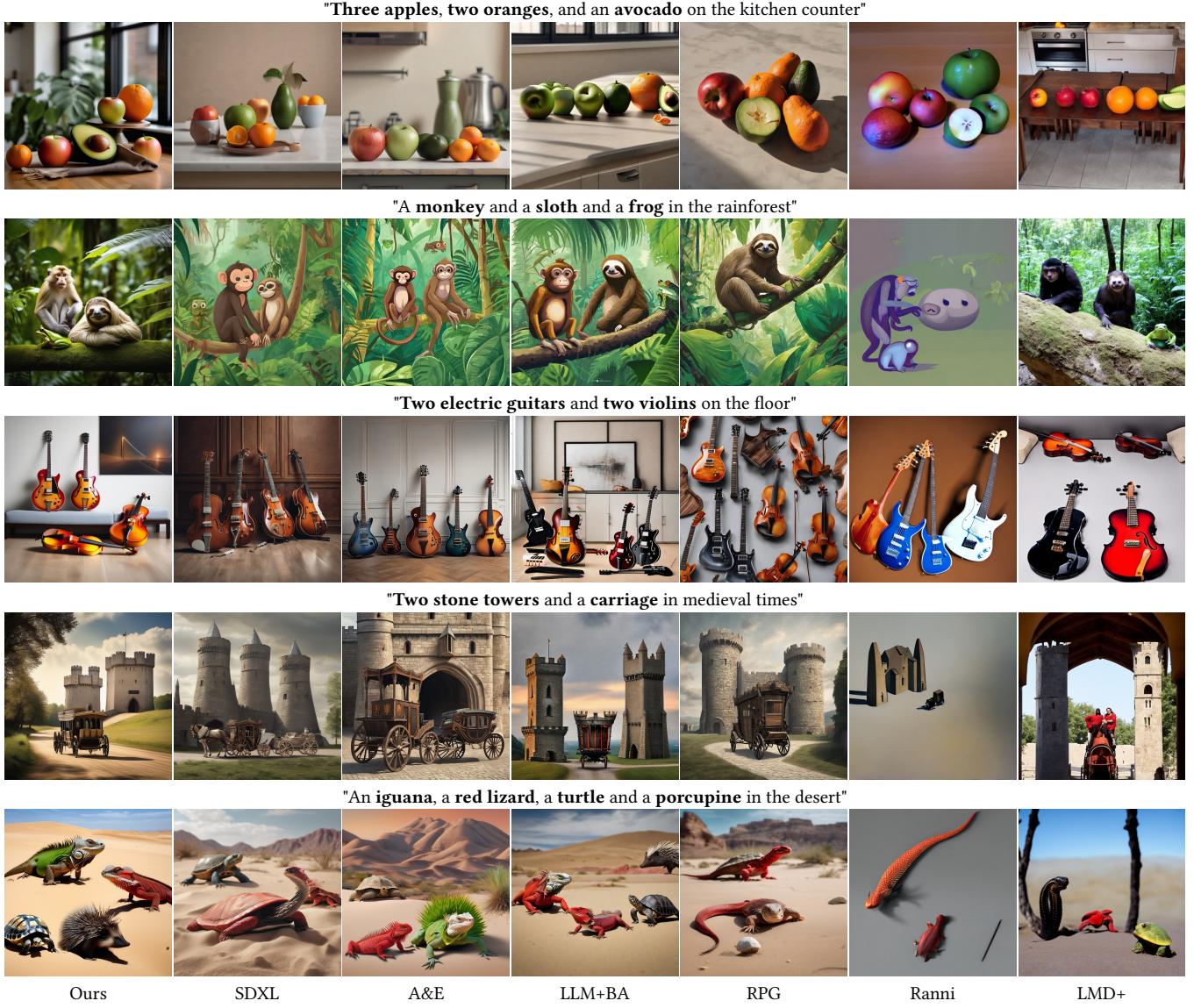


Fig. 12. Qualitative comparison of our method with baseline methods.

due to the instability of the hard-layouts during generation. Omitting $\mathcal{L}_{\text{cross}}$ or \mathcal{L}_{var} causes clusters to fragment internally, leading them to span multiple, disconnected subject instances. Additionally, parts of a subject may be absorbed into the background, resulting in shrunken or incomplete subject regions. Omitting $\mathcal{L}_{\text{dice}}$ also promotes over-generation, as oscillating cluster boundaries across timesteps lead to the emergence of redundant subjects with mixed appearances at cluster edges.

Finally, we ablate the choice of computing our $\mathcal{L}_{\text{decisive}}$ loss between the intermediate soft-layout S^{t-1} and the previous hard-layout M^t . Instead, we apply an additional denoising step on z_{t-1} before optimizing it using guidance. That denoising step is employed to extract the denoising model’s features and compute an updated

hard-layout M^{t-1} . Then, during guidance, we compute $\mathcal{L}_{\text{decisive}}$ between S^{t-1} and M^{t-1} . As evident in the second-to-right column, this approach also compromises accurate subject generation, yielding redundant subject instance due to clustering inconsistencies between timesteps.

B.4 Limitations

In Figure 14, we present two limitations of our method. First, in cluttered scenes, subjects may appear with irregular sizes or exhibit poor interaction with the background (left image). We observe that this issue also occurs with vanilla SDXL and can often be mitigated by increasing the number of denoising steps.

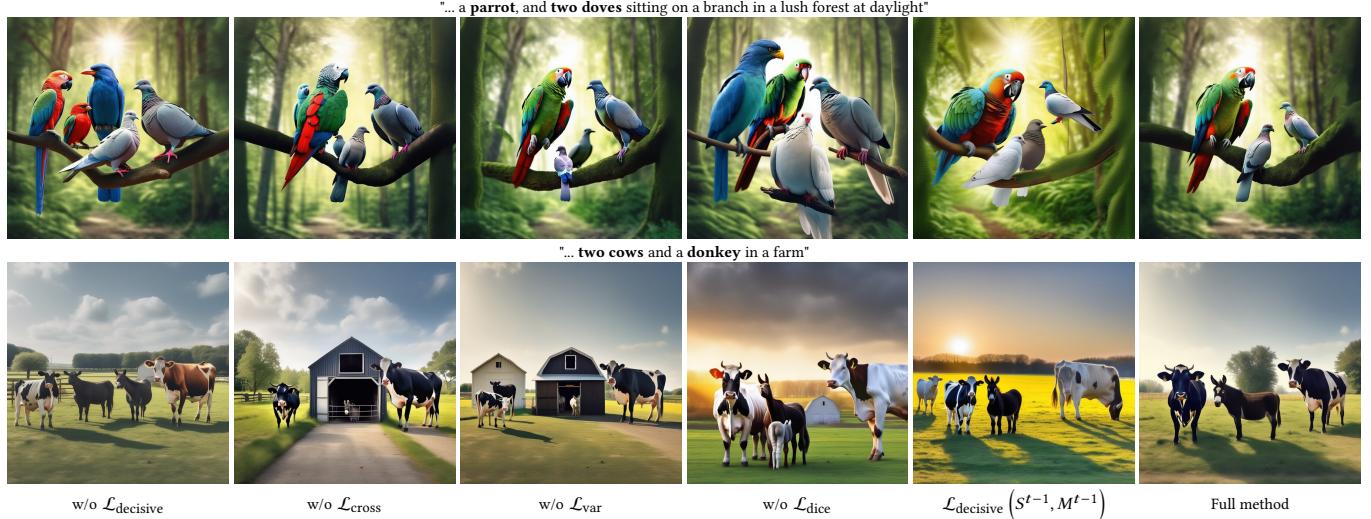


Fig. 13. Qualitative ablation. We ablate our method by skipping the guidance steps (w/o $\mathcal{L}_{\text{decisive}}$), dropping a loss term when optimizing (w/o $\mathcal{L}_{\text{cross}}$, w/o \mathcal{L}_{var} , w/o $\mathcal{L}_{\text{dice}}$), and performing an alternative guidance step ($\mathcal{L}_{\text{decisive}}(S^{t-1}, M^{t-1})$), where the loss is computed between the soft- and hard-layouts of the same timestep (instead of $\mathcal{L}_{\text{decisive}}(S^{t-1}, M^t)$). All images in each row are generated using the same seed.



Fig. 14. Limitations.

Second, since the layouts are derived from the model’s prior — which lacks a robust understanding of spatial relationships [Chatterjee et al. 2024] — subjects may sometimes fail to respect spatial constraints specified in the prompt (right image).