

MALeR: Improving Compositional Fidelity in Layout-Guided Generation

SHIVANK SAXENA, CVIT, IIIT Hyderabad, India

DHRUV SRIVASTAVA, CVIT, IIIT Hyderabad and Adobe Research, India

MAKARAND TAPASWI, CVIT, IIIT Hyderabad, India



Fig. 1. We present typical challenges of modern layout-guided text-to-image generation methods. From left-to-right, we first present the layout-guidance prompt. Next, we present some challenges: (a) *Background semantic leakage* where additional subjects appear outside the intended region; (b) *Out-of-distribution image generation* with cracked images and/or erroneous textures; and (c) *Incorrect attribute binding* with too many subjects. (d) MALeR, our approach, is able to solve these challenges and generate an accurate multi-subject multi-attribute image.

Recent advances in text-to-image models have enabled a new era of creative and controllable image generation. However, generating compositional scenes with multiple subjects and attributes remains a significant challenge. To enhance user control over subject placement, several layout-guided methods have been proposed. However, these methods face numerous challenges, particularly in compositional scenes. Unintended subjects often appear outside the layouts, generated images can be out-of-distribution and contain unnatural artifacts, or attributes bleed across subjects, leading to incorrect visual outputs. In this work, we propose MALeR, a method that addresses each of these challenges. Given a text prompt and corresponding layouts, our method prevents subjects from appearing outside the given layouts while being in-distribution. Additionally, we propose a masked, attribute-aware binding mechanism that prevents attribute leakage, enabling accurate rendering of subjects with multiple attributes, even in complex compositional scenes. Qualitative and quantitative evaluation demonstrates that our method achieves superior performance in compositional accuracy, generation consistency, and attribute binding compared to previous work. MALeR is particularly adept at generating images of scenes with multiple subjects and multiple attributes per subject.

CCS Concepts: • **Computing methodologies** → **Image manipulation**; **Computer vision**.

Authors' Contact Information: Shivank Saxena, CVIT, IIIT Hyderabad, India, shivank.saxena@research.iiit.ac.in; Dhruv Srivastava, CVIT, IIIT Hyderabad and Adobe Research, India, dsrivastava@adobe.com; Makarand Tapaswi, CVIT, IIIT Hyderabad, India, makarand.tapaswi@iiit.ac.in.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM 1557-7368/2025/12-ART236 <https://doi.org/10.1145/3763341>

Additional Key Words and Phrases: Layout-guided text-to-image generation, Subject-attribute binding, Diffusion models, Latent optimization

ACM Reference Format:

Shivank Saxena, Dhruv Srivastava, and Makarand Tapaswi. 2025. MALeR: Improving Compositional Fidelity in Layout-Guided Generation. *ACM Trans. Graph.* 44, 6, Article 236 (December 2025), 12 pages. <https://doi.org/10.1145/3763341>

1 Introduction

Diffusion models have achieved remarkable success in generating high quality and realistic images from text prompts [Dhariwal and Nichol 2021; Esser et al. 2024, 2021; Ho et al. 2020; Kingma et al. 2021; Nichol and Dhariwal 2021; Podell et al. 2024; Rombach et al. 2022; Saharia et al. 2022; Song et al. 2021b]. These models often use techniques such as classifier-free guidance [Ho and Salimans 2021] for text-conditioned image generation. However, models struggle with complex text prompts involving multiple subjects and attributes [Chefer et al. 2023; Dahary et al. 2024; Rassin et al. 2023]. Common failure modes include catastrophic neglect (skipped subjects), creation of extra subjects (e.g. two dogs when asked to generate one), and incorrect subject-attribute associations.

To address catastrophic neglect and improve prompt adherence, Chefer et al. [2023] introduced a cross-attention-guided excitation mechanism that optimizes the latent state of the diffusion model during inference. They maximize the cross-attention between subject tokens and image patches, ensuring that each subject in the prompt exerts sufficient influence on the generated image. Follow-up works (e.g. [Agarwal et al. 2023; Guo et al. 2024; Li et al. 2023a]) adopt similar latent optimization paradigms, often using variants of cross-attention-based losses. However, the optimization process can drive the latents out-of-distribution resulting in incoherent generations. Thus, generating multiple samples with different random initializations is common to achieve the desired scene composition.

To enhance user control, recent works condition text-to-image models through the use of visual layouts to indicate spatial location of subjects. Grounding inputs such as bounding boxes [Li et al. 2023b], depth maps [Huang et al. 2023; Zhang et al. 2023], semantic maps, or scribbles [Huang et al. 2023; Lv et al. 2024; Wang et al. 2024; Zhang et al. 2023], improve control and guide the generation process. While semantic and depth maps provide fine-grained spatial information, they are often impractical for users to manually construct or edit. On the other hand, bounding boxes offer a simple and intuitive alternative to specify desired subject locations.

However, such layout-guided methods inherit the same fundamental limitations observed in text-to-image generation. They are susceptible to distributional drift due to latent optimization, and additionally face the challenge of ensuring adherence to the provided layout. While some methods attempt to mitigate these issues by manipulation of cross-attention maps during the sampling process [Chefer et al. 2023; Chen et al. 2024; Feng et al. 2023], recent approaches [Dahary et al. 2024; Phung et al. 2024] also incorporate self-attention maps into the guidance process. To some extent, this helps, but creates new challenges. For example, layout-guided methods suffer a problem of *background semantic leakage*, where unintended (additional) subjects appear outside or near the specified bounding boxes (Fig. 1a). Moreover, the constraints also make the generations brittle resulting in *out-of-distribution images* characterized by texture-less subjects, tiling or cracking, and other unnatural artifacts (Fig. 1b). Interestingly, while some random state initializations achieve close to desirable compositions, many lead to failures—necessitating tens of attempts to obtain a good image.

Additionally, crafting diverse compositional scenes requires precise attribute binding between subjects. Prior works [Feng et al. 2023; Jiang et al. 2024; Li et al. 2023a; Meral et al. 2024; Rassin et al. 2023] associate attributes with their corresponding subjects by leveraging cross-attention-based losses or similarity measures. However, these methods are limited to handling one or two subjects and often struggle in complex compositions resulting in *attribute leakage* across subjects and background regions, or *subject blending*, where individual identities become visually entangled (Fig. 1c). This highlights the need for a training-free layout-guided generation framework that not only enforces spatial alignment, but also supports robust attribute binding across multiple subjects—particularly in complex, multi-attribute scenes.

In this work, we propose Masked Attribute-aware Latent Regularization (MALeR, *n. painter* in German) to address all three challenges. We introduce a *masked latent regularization* strategy to prevent background semantic leakage during latent optimization. Specifically, we discourage the emergence of subject-like patterns outside the designated bounding boxes by anchoring background latents close to their original values. Second, we perform *in-distribution latent alignment* to prevent out-of-distribution images during latent optimization. More concretely, during early denoising steps, we encourage the optimized latents to remain close to the prior Gaussian distribution through an alignment term based on KL-divergence. Third, we propose a novel *subject-attribute association* loss that encourages similarity (dissimilarity) between masked regions of cross-attention maps of paired (unpaired) nouns and adjectives. This formulation not only enables accurate attribute binding across multiple subjects,

but also supports *multiple attributes* to be associated with each subject, enabling generation of rich and compositional scenes with precise subject-attribute association.

The main contributions of our work are summarized next. (i) We identify *background semantic leakage* as a limitation of current layout-guided generation methods and propose *masked latent regularization* as a way to address it. (ii) To prevent out-of-distribution artifacts, we regularize the latents through *in-distribution latent alignment*. (iii) We introduce a novel *subject-attribute association loss* to ensure correct binding of *multiple attributes* in compositional generation. (iv) Thorough experiments are presented using the same random seeds. We qualitatively show that MALeR succeeds at generating images for difficult prompts containing multiple attributes. We also present quantitative comparisons against previous works on DrawBench [Saharia et al. 2022] and HRS [Bakr et al. 2023] benchmarks and establish a new state-of-the-art performance.

2 Related Work

Text-to-image (T2I) models have improved a lot [Balaji et al. 2022; Podell et al. 2024; Ramesh et al. 2021; Rombach et al. 2022; Saharia et al. 2022; Sauer et al. 2024]. Driven by the success of Transformers [Dosovitskiy et al. 2021; Vaswani et al. 2017], we see the emergence of Transformer-based T2I models [Esser et al. 2024; Gao et al. 2023; Peebles and Xie 2023; Zheng et al. 2024]. However, despite the tremendous success, generating images aligned with compositional, multi-subject text prompts remains a challenge.

Controllable generation. To improve T2I model controllability, techniques such as prompt optimization [Hao et al. 2023; Hertz et al. 2022; Mo et al. 2024; Witteveen and Andrews 2022], reward based tuning [Xu et al. 2023], or inference-time latent updates [Chefer et al. 2023] are popular. Subsequent works build upon latent update techniques by manipulating cross- and self-attention maps [Battash et al. 2024; Feng et al. 2023; Guo et al. 2024; Li et al. 2023a; Sundaram et al. 2024; Tang et al. 2023; Wu et al. 2023] during inference to minimize catastrophic neglect. However, these methods may produce out-of-distribution and unnatural images due to latent state optimization during inference. In addition to neglect, some T2I methods aim to address incorrect attribute binding [Feng et al. 2023, 2024; Jiang et al. 2024; Li et al. 2023a; Meral et al. 2024; Rassin et al. 2023], but are often limited to few subjects and struggle in compositional scenes with multiple subjects involving multiple attributes. We propose MALeR, a training-free T2I approach that addresses the above challenges of catastrophic neglect, out-of-distribution images, and incorrect attribute binding.

Layout-guided control with boxes. Several layout-guided T2I methods either train or fine-tune models or external modules [Avrahami et al. 2023; Feng et al. 2024; Gu et al. 2025; Li et al. 2023b; Nichol et al. 2022; Nie et al. 2024; Qu et al. 2023; Wang et al. 2024; Wu et al. 2024; Yang et al. 2023; Zhang et al. 2023; Zheng et al. 2023; Zhou et al. 2024] for layout-guided generation. However, these methods require extensive computational resources. As an alternative, several training-free methods have emerged [Balaji et al. 2022; Bansal et al. 2023; Bar-Tal et al. 2023; Chen et al. 2024; Dahary et al. 2024; Endo 2024; Phung et al. 2024; Xiao et al. 2024; Xie et al. 2023; Zhao et al. 2023]. Among them, methods such as Layout-guidance [Chen et al.

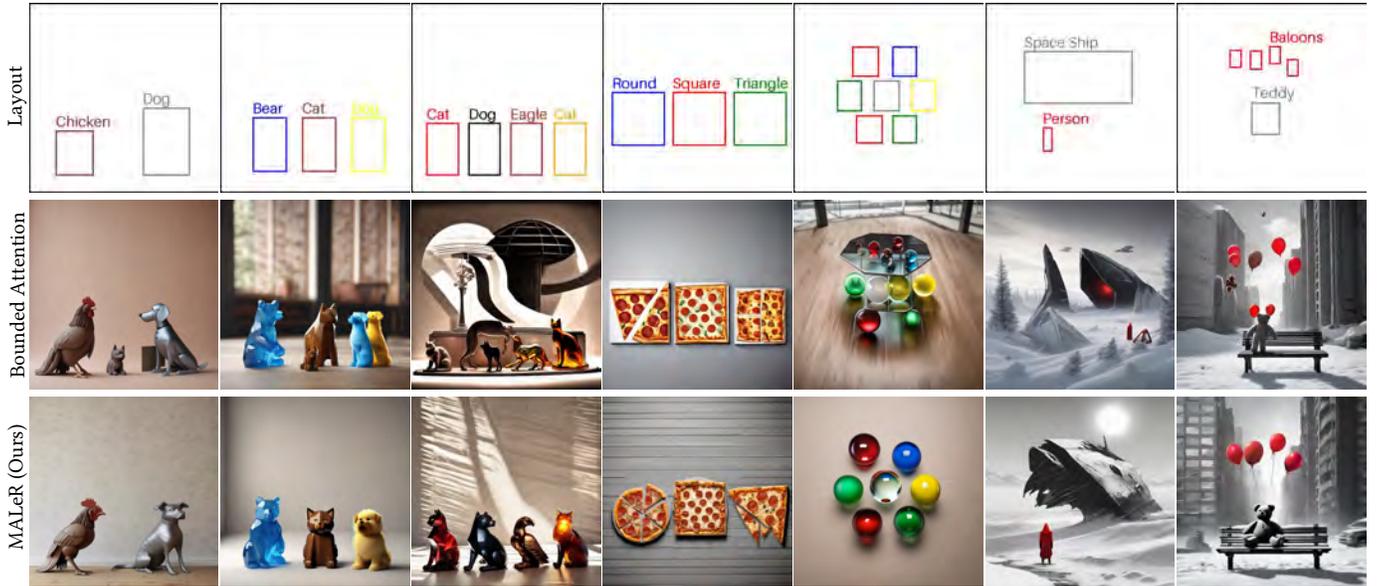


Fig. 2. MAlEr (SDXL) outperforms Bounded Attention [Dahary et al. 2024] on complex prompts with multiple subjects and attributes. Both BA and MAlEr use the same seed. The prompts from L2R are: 1. A realistic photo of a **brown wooden chicken** and a **gray metallic dog**. 2. A realistic photo of a **blue crystal bear** and a **brown wooden cat** and a **yellow fluffy dog**. 3. A realistic photo of a **shiny red crystal cat** and a **black matte plastic dog** and a **rustic bronze eagle** and a **glowing amber cat**. 4. A **round pizza** and a **square pizza** and a **triangle pizza**. 5. A realistic photo of **two red glass sphere** and a **blue glass sphere** and **two green glass sphere** and a **yellow glass sphere** and a **white glass sphere**. 6. A black and white concept art of a **crashed spaceship** partially buried in icy landscape and a **red hooded person** is watching it from a distance. 7. A black and white concept art of a **destroyed apocalyptic city** covered with snow and a **decaying teddy bear** on a bench with **four red balloons** tied.

2024], R&B [Xiao et al. 2024], or BoxDiff [Xie et al. 2023] use the cross-attention map to enable layout guidance. Others like Attention Refocusing [Phung et al. 2024] and Bounded Attention [Dahary et al. 2024] use cross- and self-attention maps for layout guidance through latent state optimization. However, for multi-subject prompts, we observe that such methods exhibit background semantic leakage, generate out-of-distribution images, and spread attributes across subjects. Our approach addresses these challenges through regularization of the latent updates and promoting correct subject-attribute pairs while demoting others.

3 Our Approach: MAlEr

We present the components of Masked Attribute-aware Latent Regularization (MAlEr): (i) masked latent regularization prevents background semantic leakage (Sec. 3.2), (ii) in-distribution latent alignment avoids artifacts (Sec. 3.3), and (iii) subject-attribute association improves binding (Sec. 3.4). First, we re-visit the fundamentals of inference-time latent optimization in layout-guided T2I models.

3.1 Preliminaries

Different from pixel space diffusion [Ho et al. 2020], a Latent Diffusion Model (LDM) [Rombach et al. 2022] operates in the latent embedding space. It has an autoencoder that encodes an image x to the latent space $z = \mathcal{E}(x)$ and reconstructs the image through a decoder $\hat{x} = \mathcal{D}(z)$. During training, noise is gradually added to the original latent state z_0 to obtain z_t . During inference, a UNet [Ronneberger et al. 2015], equipped with self- and cross-attention layers

acts as a denoiser. Starting from random noise $z_T \sim \mathcal{N}(0, I)$, the denoiser estimates and removes the noise $\hat{\epsilon}_t = \phi(z_t, t, y)$ conditioned on the time step t and the text prompt y . Specifically, we adopt the DDIM [Song et al. 2021a] update mechanism.

Attention layers. At each layer of the UNet, the prompt embedding y is injected through cross-attention layers to steer the image towards the prompt. Specifically, spatial UNet features $\phi(z_t)$ are projected to obtain queries $Q = W_q\phi(z_t)$, while keys $K = W_k y$ and values $V = W_v y$ are obtained using the prompt embedding. The cross-attention map at t is calculated as $A_t^c = \text{softmax}(QK^T)$. While the cross-attention maps $A_t^c \in \mathbb{R}^{hw \times n}$ capture the relationship between the latent of spatial dimensions hw against n prompt tokens, the self-attention maps $A_t^s \in \mathbb{R}^{hw \times hw}$ depict the spatial correspondence between latents.

Latent optimization. A popular approach to improve inference-time prompt adherence of LDMs is to optimize the latent z_t at each time step during sampling:

$$z'_t \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}, \quad (1)$$

where α_t is the update rate. \mathcal{L} defines the desired objective, e.g. Chefer et al. [2023] minimize catastrophic neglect by ensuring high attention of at least one latent region to each subject token $s_i \in \mathcal{S}$:

$$\mathcal{L} = \max_{s_i} (1 - \max_{hw} (A_t^c(s_i))), \quad (2)$$

where $A_t^c(s_i) \in \mathbb{R}^{hw}$ represents the spatial cross-attention to token s_i of the prompt.

Layout-guided generation. Subjects in the prompt are associated with spatial guidance in the form of bounding boxes by pairing token indices with boxes (s_i, b_i) . Latent optimization encourages subjects to be present inside the bounding boxes [Dahary et al. 2024; Phung et al. 2024]. Specifically, the latent updates minimize two losses: a cross-attention loss \mathcal{L}_c , encourages generated subjects to be present inside the corresponding bounding boxes; and a self-attention loss \mathcal{L}_s , prevents latent pixels from attending to irrelevant regions. For both losses, previous works [Dahary et al. 2024; Phung et al. 2024; Xiao et al. 2024] use an intersection-over-union (IoU) formulation that encourages attention to focus inside the bounding box region while disregarding the rest. For a subject token index s_i with box b_i , this is defined as:

$$\mathcal{L}_i = 1 - \frac{\sum \hat{A}_t[b_i]}{\sum \hat{A}_t[b_i] + \gamma \sum \hat{A}_t[\bar{b}_i]}, \quad \text{and} \quad \mathcal{L}_{\text{iou}} = \sum_i \mathcal{L}_i^2, \quad (3)$$

where \hat{A}_t is the aggregated self- or cross-attention map (heads and layers) at step t . $\hat{A}_t[b_i]$ corresponds to attention within the subject's box b_i and $\hat{A}_t[\bar{b}_i]$ to regions outside the box. γ is the number of subjects in the prompt and it amplifies the attention towards the background. For details, we refer the reader to [Dahary et al. 2024] and adopt this IoU loss as our baseline.

3.2 Masked Latent Regularization

Layout-guided generation methods aim to create images where subjects adhere to both, the text prompts and the bounding boxes. However, we observe that current methods exhibit *background semantic leakage* where multiple subjects, not specified in the prompts (box, text, or both), appear in the image (see Fig. 1a). We see that this becomes frequent with increasing number of subjects.

The first few inference time steps of diffusion models largely determine the layout [Hertz et al. 2022]. Subjects emerge first while fine-grained details later. While previous works optimize the latent for 15 to 25 steps [Dahary et al. 2024], we find that the object layout is already determined in the first 5 steps. In fact, over optimization of the latent leads to out-of-distribution images.

While optimizing the latent z_t , we wish to discourage subject patterns from forming outside bounding boxes. We do so by constraining the background latents to remain close to their original values. Consider z_t^{ref} as a detached copy of the latent before the update (z_t in Eq. (1)). We create a binary mask M with value 1 in regions corresponding to a bounding box and 0 elsewhere, and let \bar{M} represent the inverted background mask. We introduce a masked regularization term that penalizes deviations and incorporate it in the latent update as:

$$\mathcal{L}_{\text{mask}} = \left\| (z_t^{(k)} - z_t^{\text{ref}}) \odot \bar{M} \right\|_1, \quad \text{and} \quad (4)$$

$$z_t^{(k+1)} \leftarrow z_t^{(k)} - \alpha_t \cdot \nabla_{z_t} (\mathcal{L}_{\text{iou}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}), \quad (5)$$

where λ_{mask} is a hyperparameter and we perform iterative refine-ment k times at each time step t , similar to [Chefer et al. 2023]. This approach is illustrated in Fig. 3a with highlighted masks.

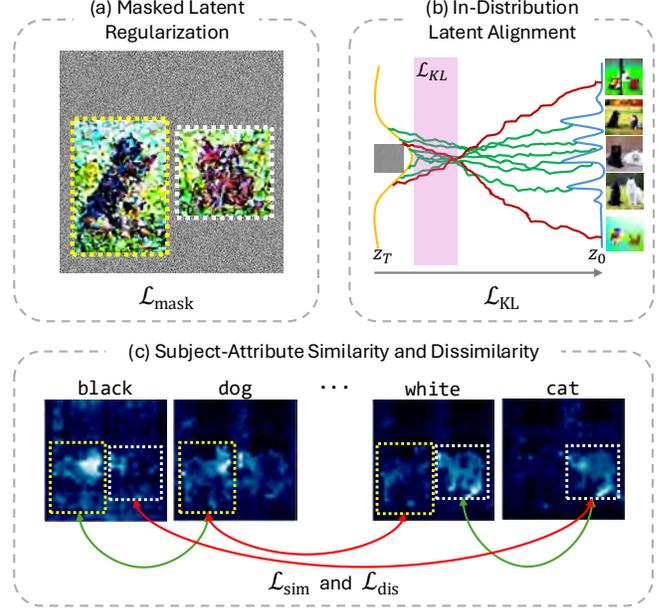


Fig. 3. **MALeR overview.** We illustrate three key components for layout-guided compositional scene generation. (a) Masked latent regularization prevents background semantic leakage, (b) KL-based alignment keeps the latent in-distribution during optimization, and (c) layout-guided subject-attribute association enables accurate compositional binding.

3.3 In-Distribution Latent Alignment

Latent space updates risk pushing z_t away from its typical noise distribution, resulting in out-of-distribution images. In layout-guided methods, the imbalance between latent updates within and outside the boxes further increases this chance. To mitigate this problem, we align the latent back to a stable prior distribution. In the early stages of the denoising process, the latent is close to the random noise distribution, $z_t \sim \mathcal{N}(0, \mathbf{I})$. Thus, we introduce an alignment term (for 5 steps) based on the Kullback-Leibler (KL) divergence [Kullback and Leibler 1951] and add it to the latent update loss as:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\mathcal{N}(\mu_{z_t}, \sigma_{z_t}) \parallel \mathcal{N}(0, 1)), \quad (6)$$

$$\mathcal{L} = \mathcal{L}_{\text{iou}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}. \quad (7)$$

λ_{KL} is a hyperparameter that controls the strength of the KL term. The combined formulation not only prevents background leakage but also constrains the tendency of \mathcal{L}_{iou} and $\mathcal{L}_{\text{mask}}$ from pushing z_t out-of-distribution during the early phase of optimization (see Fig. 3b). The outcome is generation of higher fidelity images without needing multiple random initializations (Fig. 2, Fig. 4).

3.4 Layout-guided Subject-Attribute Association

A challenge in layout-guided scene composition is binding attributes to the right subjects. To address this, we extract nouns (subjects) and their modifiers (attributes) from the prompt and propose constraints on the attention mechanism. Specifically, we propose that the spatial cross-attention patterns should be similar for aligned subject-attribute pairs *and* dissimilar for unaligned pairs.

Concretely, consider a set of subject tokens indicated through their indices $\mathcal{S} = \{s_1, \dots, s_S\}$ in the prompt and associated with boxes $\mathcal{B} = \{b_1, \dots, b_S\}$. Let us denote the set of attribute token indices for each subject as $\mathcal{A} = \{a_1, \dots, a_S\}$ (no attribute is subsumed with $a_i = \emptyset$). In addition to the subject token and bounding box (s_i, b_i) , we now include attribute token indices (s_i, b_i, a_i) . Note, a_i may correspond to multiple attribute tokens depending on the prompt (e.g. white and marble lion in Fig. 1). Recall, $A_t^c \in \mathbb{R}^{h \times w \times n}$ represents the cross-attention map at time step t . For a specific subject token s_i , we denote $A_t^c(s_i)[b_i] \in \mathbb{R}^{h_i \times w_i}$ as the cross-attention in the spatial region corresponding to the bounding box b_i of size $h_i \times w_i$. Similarly $A_t^c(a_i)[b_i]$ corresponds to the attention scores of the box region with respect to the attribute token(s). We renormalize these attention patches to be a probability.

Subject-attribute similarity. The association between paired subjects and attributes is improved by encouraging similar cross-attention maps within the bounding box. We calculate the similarity loss between subject s_i and attribute a_i using a symmetric KL divergence:

$$\mathcal{L}_{\text{sim}}(i) = D_{\text{sym}}(A_t^c(s_i)[b_i], A_t^c(a_i)[b_i]), \quad \text{where} \quad (8)$$

$$D_{\text{sym}}(P, Q) = \frac{1}{2}D_{\text{KL}}(P\|Q) + \frac{1}{2}D_{\text{KL}}(Q\|P). \quad (9)$$

The total similarity loss $\mathcal{L}_{\text{sim}} = \text{mean}_i \mathcal{L}_{\text{sim}}(i)$.

Subject \times attribute dissimilarity. Minimizing the similarity loss alone is insufficient to bind attributes to objects (Fig. 6e). Similar to the triplet loss [Schroff et al. 2015], we propose a dissimilarity loss between unaligned subject-attribute boxes. We consider all mismatched subject-attribute combinations (s_i, b_i, a_j) , $j \neq i$ for whom the cross-attention maps should be dissimilar. The dissimilarity loss is formulated as a negative symmetric KL divergence operating on the cross-attention region b_i for tokens s_i and a_j :

$$\mathcal{L}_{\text{dis}}(i, j) = -D_{\text{sym}}(A_t^c(s_i)[b_i], A_t^c(a_j)[b_i]). \quad (10)$$

The total loss is $\mathcal{L}_{\text{dis}} = \text{mean}_{i,j,i \neq j} \mathcal{L}_{\text{dis}}(i, j)$.

Why adopt symmetric KL divergence? Cross-attention maps have been treated as a probability distribution in prior works, and the distance between subject and attribute cross-attention maps is minimized to improve binding. Li et al. [2023a] employ JS-Divergence, while Jiang et al. [2024] use symmetric KL divergence. However, we observe that directly minimizing the distance between cross-attention maps is ineffective in the presence of multiple subjects with multiple attributes. Layout conditioning in our method specifies the exact regions where the subject and attribute cross-attention maps need to be similar. Thus, we minimize the distance between *normalized cross-attention maps* within the masked regions corresponding to each subject and its attributes. However, even after this alignment, attributes may still leak to other subjects in multi-subject scenes (Fig. 6e). Our masked dissimilarity loss prevents this leakage by maximizing the distance between each subject and its non-corresponding attributes. Finally, we empirically find symmetric KL to be much more effective than JS-divergence.

Total attribute loss and visualization. The final attribute association loss, $\mathcal{L}_{\text{att}} = \lambda_{\text{sim}} \mathcal{L}_{\text{sim}} + \lambda_{\text{dis}} \mathcal{L}_{\text{dis}}$, where λ_{sim} and λ_{dis} are hyperparameters, controlling the strength of similarity and dissimilarity. Fig. 3c shows cross-attention maps for the example prompt “a black

dog and a white cat” with $(s_1: \text{dog}, a_1: \text{black})$ and $(s_2: \text{cat}, a_2: \text{white})$. The cross-attention maps for aligned pairs (dog, black) or (cat, white) show high activations within the provided box guidance. Similarly, the attention maps for unaligned pairs (cat, black) and (dog, white) show significant differences.

Final training objective. The overall objective is a combination of multiple terms inducing regularization, alignment to the prior distribution, and encouraging correct subject-attribute association:

$$\mathcal{L} = \mathcal{L}_{\text{iou}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{att}} \mathcal{L}_{\text{att}}. \quad (11)$$

This combined formulation improves background semantic leakage, reduces out-of-distribution images, and provides more accurate attribute binding even with multiple attributes for each subject, generating high fidelity compositional scenes.

4 Experiments

Baselines. We compare MALeR against seven previous approaches in layout-guided T2I. They include: GLIGEN [Li et al. 2023b], Attention Refocusing [Phung et al. 2024], BoxDiff [Xie et al. 2023], Layout-Guidance [Chen et al. 2024], ReCo [Yang et al. 2023], R&B [Xiao et al. 2024], and Bounded Attention (BA) [Dahary et al. 2024] (equivalent of \mathcal{L}_{iou} only loss). We present results with two backbones: SD-1.5 and SDXL [Podell et al. 2024], for fair comparisons.

Benchmarks. We evaluate layout-guided T2I methods on two benchmarks: DrawBench [Saharia et al. 2022] and HRS [Bakr et al. 2023]. DrawBench is well-established, with challenging prompts to evaluate spatial reasoning and counting capabilities of T2I models. It provides 39 prompts for *counting* (19) and *spatial* (20) relationships. We also evaluate MALeR’s ability to bind attributes correctly using 25 prompts featuring 9 colors from the *color* task. Notably, prior methods have skipped this category. On HRS benchmark, we follow the protocol established by R&B [Xiao et al. 2024], and report results on spatial, color, and size, to demonstrate the effectiveness of our approach. We use bounding boxes provided by Phung et al. [2024], that are generated automatically using GPT-4.

Evaluation metrics. We follow standard evaluation protocols [Dahary et al. 2024; Phung et al. 2024; Xiao et al. 2024]. For counting, we use an off-the-shelf object detector and compare its output to the ground-truth prompt and calculate precision, recall, and F1 score. Accuracy is adopted for the spatial, color, and size prompts.

Implementation details. All experiments are performed on A6000 GPUs. On DrawBench, results are averaged across 4 seeds (0, 42, 2718, 31415). For HRS, we adopt seed 0. All qualitative comparisons are made across the same seed. We empirically choose $\lambda_{\text{mask}}=0.01$, $\lambda_{\text{KL}}=5$, and $\lambda_{\text{sim}}=\lambda_{\text{dis}}=\lambda_{\text{att}}=1$ as they give consistently good results. $\mathcal{L}_{\text{mask}}$ and \mathcal{L}_{KL} are applied for the first 5 denoising steps, while \mathcal{L}_{att} is applied for the first 18 of 50 denoising steps. At each denoising step when the loss term is applied we perform $k=5$ gradient descent iterations. For the step size α in Eq. (1), we linearly decrease it from 30 to 8 across the denoising steps and the attention map layers used for optimization are the same as BA [Dahary et al. 2024].

4.1 Results

Qualitative comparison. Fig. 2 demonstrates MALeR’s ability to create compositional scenes involving multiple subjects and multiple

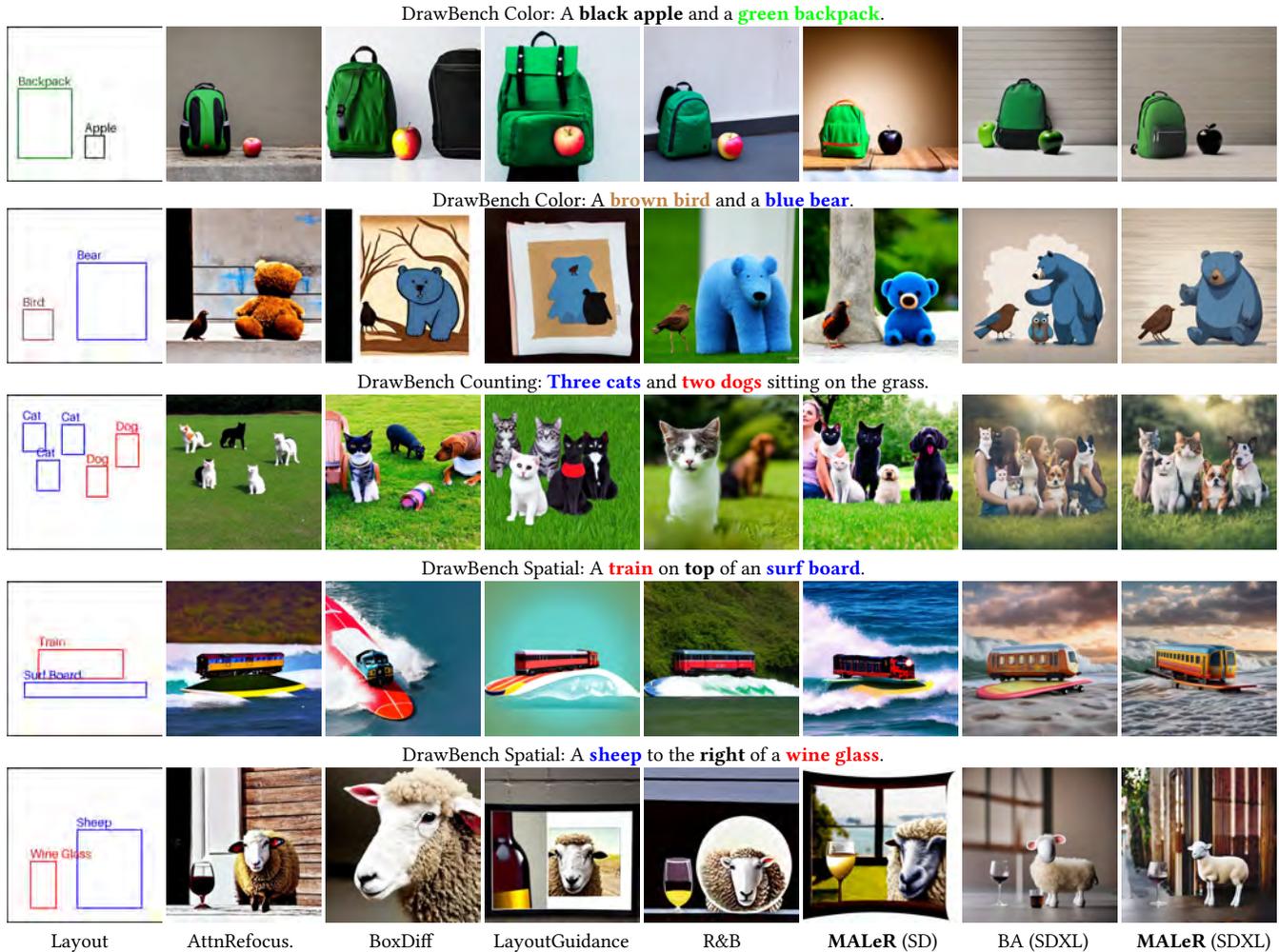


Fig. 4. We compare generated images on DrawBench. Each row uses the same random seed. The text prompt is shown above and layout in column 1. We compare MALeR (SD) against: Attention Refocusing [Phung et al. 2024], BoxDiff [Xie et al. 2023], Layout Guidance [Chen et al. 2024], and R&B [Xiao et al. 2024]. Bounded Attention (BA) [Dahary et al. 2024] with SDXL is compared against MALeR (SDXL). MALeR shows strong adherence to the prompt (subjects, attributes, and layout), generates high quality images without background semantic leakage, and correctly localizes the subjects.

attribute types. Our approach performs well beyond the typical color attribute and showcases subjects with various material properties. For example, in columns 1-3, our method accurately applies color and material properties to each subject, while BA creates extra subjects and confuses attributes across them. Columns 4-7 further highlight our model’s ability to create images with correct pizza shapes, location and color for multiple subjects (spheres), and rich art-like scenes (concept art) that adhere to the layout guidance. The spheres image that requires creation of 7 subjects is particularly challenging for the baseline (BA) resulting in erroneous number of objects and blended colors.

Next, we compare generated images on DrawBench in Fig. 4. MALeR produces outputs with adherence to the prompt and layout as compared to other works. The first example (black apple, green backpack) is notable as most methods generate a red apple while BA shows some leakage in the color across subjects. The third example with 3 dogs and 2 cats is difficult and most methods get

it wrong. AttnRefocus has unnatural artifacts, BoxDiff has wrong count and location, LayoutGuidance generates questionable dogs, R&B generates only 2 subjects, and BA generates people in the background. While MALeR (SD) also generates a person, with SDXL, our image is the only correct generation. Additional qualitative results can be seen in Fig. 8 and Fig. 9. The former shows good outputs from MALeR for multiple random seeds as compared to BA, while the latter presents several examples of complex prompts and layouts. We also show generated images by varying the subject attributes across two dimensions in Fig. 5.

Quantitative comparison. Tab. 2 shows that MALeR outperforms all prior methods across the various tasks on DrawBench and two of three tasks on HRS. We ensure fair comparisons and promote reproducibility by running all methods using the same random seeds. We also present results while using comparable backbones, SD-1.5 and SDXL. For spatial prompts, MALeR improves over the strongest



Fig. 5. **Color variation** across the wizard and the thunderbolt for the prompt: *A concept art of an icy landscape with a [red | black] robe wizard summoning a [pink | green | purple | blue] colored magic thunderbolt from air.* All images are generated with the same random seed.

baseline (GLIGEN) by +6.0% on DrawBench and R&B by +3.7% on HRS. A similar large improvement is seen in color prompts with +17% improvement on DrawBench and +3.8% on HRS. These gains highlight the effectiveness of the subject-attribute association losses in generating more accurate associations. On the DrawBench counting task, our method achieves the highest F1 score (0.88), matching the previous best baseline (R&B) while outperforming others. While MAlER performs well on spatial, color, and counting metrics, we observe lower performance on the HRS size task, partly owing to inaccurate and confusing aspect ratios of guiding boxes.

Notably, MAlER consistently outperforms Bounded Attention, our closest baseline, that uses the same SDXL backbone by a significant margin. With SD-1.5 on DrawBench, MAlER (vs. BA) achieves 0.78 (+10%) on spatial accuracy, 0.42 (+8%) on color accuracy, and a comparable F1 score of 0.85 (+1%) on counting. Similarly, on HRS with SD-1.5, MAlER reaches 31.8% (+0.9%) on spatial accuracy and 40.4% (+7.4%) on color accuracy.

FID scores. We compare perceptual quality of generated images on Drawbench for Vanilla SDXL, BA, and MAlER. Tab. 1 shows that all three models achieve comparable scores on perceptual fidelity. In fact, MAlER outperforms BA slightly and is comparable to SDXL.

4.2 User Studies

We conduct two user studies to further validate and compare images generated by BA and MAlER.

Average human ranking study. In the first study, we select all four challenging prompts visualized in Fig. 8 and generate outputs using MAlER and BA for 10 random seeds (0-9). Five independent (non-author) raters score each of the 80 images (4 prompts, 10 seeds, 2 methods) on a Likert scale of 1-5, yielding a total of 400 ratings. The mean scores in Tab. 3a show that MAlER consistently outperforms BA across all prompts. Further, the standard deviations provide interesting insights. For prompts 1, 3, and 4, BA shows low mean and low standard deviation, indicating poor outputs across most seeds. For prompt 2, MAlER shows high mean with low standard deviation, indicating consistently strong outputs across the seeds.

Table 1. FID scores on DrawBench for Vanilla SDXL, BA, and MAlER.

	SDXL	BA	MAlER
FID (\downarrow)	161.03	164.59	163.02

Table 2. We report results on DrawBench (Spatial, Color, Counting) and HRS Benchmark (Spatial, Color, Size). Baseline acronyms are: AttnRef: Attention Refocusing, LGuidance: Layout Guidance, and BoundAttn: Bounded Attention (equivalent of only \mathcal{L}_{iou} loss). ReCo with *, fine-tunes SD. Except HRS Size (some layout challenges), MAlER achieves best performance.

Method	Base	DrawBench					HRS Benchmark		
		Spat. Acc.	Col. Acc.	Counting P	Counting R	Counting F1	Spat. Acc.	Col. Acc.	Size Acc.
GLIGEN	SD1.4	0.75	0.12	0.77	0.78	0.77	27.7	15.8	66.5
ReCo	SD1.4*	0.70	0.19	0.75	0.96	0.84	25.9	20.0	75.5
AttnRef.	SD1.4	0.74	0.31	0.80	0.79	0.79	32.0	31.3	72.5
LGuidance	SD1.5	0.65	0.31	0.83	0.75	0.79	15.9	17.4	60.5
R&B	SD1.5	0.68	0.35	0.94	0.82	0.88	34.0	34.3	77.8
BoxDiff	SD2.1	0.66	0.26	0.92	0.77	0.84	21.0	1.9	74.9
BoundAttn	SD1.5	0.68	0.34	0.86	0.82	0.84	30.9	33.0	57.5
MAlER	SD1.5	0.78	0.42	0.89	0.81	0.85	31.8	40.4	56.3
BoundAttn	SDXL	0.69	0.44	0.74	0.95	0.83	33.9	37.5	64.7
MAlER	SDXL	0.81	0.61	0.81	0.96	0.88	37.7	41.3	59.9

Table 3. User study results: We conduct user studies to compare BA and MAlER. (a) Average Human Ranking (AHR) on four challenging prompts with mean Likert scores (1–5); images shown in Fig. 8. (b) Fraction of images showing error-types: Background Semantic Leakage (BSL), Out of Distribution (OOD), Attribute Leakage (AL), and Other Errors (OE). The user studies confirm that MAlER’s outputs are better than BA.

(a) Average Human Ranking					
Method	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Avg.
BA	2.12±0.688	2.50±1.047	1.84±0.506	2.14±0.584	2.15
MAlER	3.44±0.837	3.92±0.736	2.26±0.769	2.90±0.796	3.13

(b) Error-Type Analysis				
Method	BSL	OOD	AL	OE
BA	0.445	0.305	0.590	0.325
MAlER	0.240	0.230	0.400	0.185
Relative-Improvement Δ	+46%	+25%	+32%	+43%

Error analysis. In the second user study, we analyze the kind of errors exhibited in generated images. Beyond the three main error types addressed in this work (see Fig. 1), we group all *other* observed issues into the fourth category “Other Errors”. For the same 80 images, we ask five raters to identify whether an error type is visible in each image (400 ratings). As seen in Tab. 3b, MAlER exhibits errors in fewer images as compared to BA.

4.3 Ablation Study

We perform ablation experiments for different modules of our method and show their contributions. Specifically, we assess the three new loss terms (see Eq. (11)): (i) masked latent regularization, (ii) in-distribution alignment, and (iii) subject-attribution association.

Tab. 4 presents results of including/removing different loss terms on DrawBench. As expected, applying masked regularization (\mathcal{L}_{mask}) results in a performance drop (row 2). While this term is effective at preventing background semantic leakage, it causes the latents to drift away from the prior distribution, a common issue in latent optimization methods. In contrast, incorporating the KL alignment term leads to small, but consistent improvements across all DrawBench

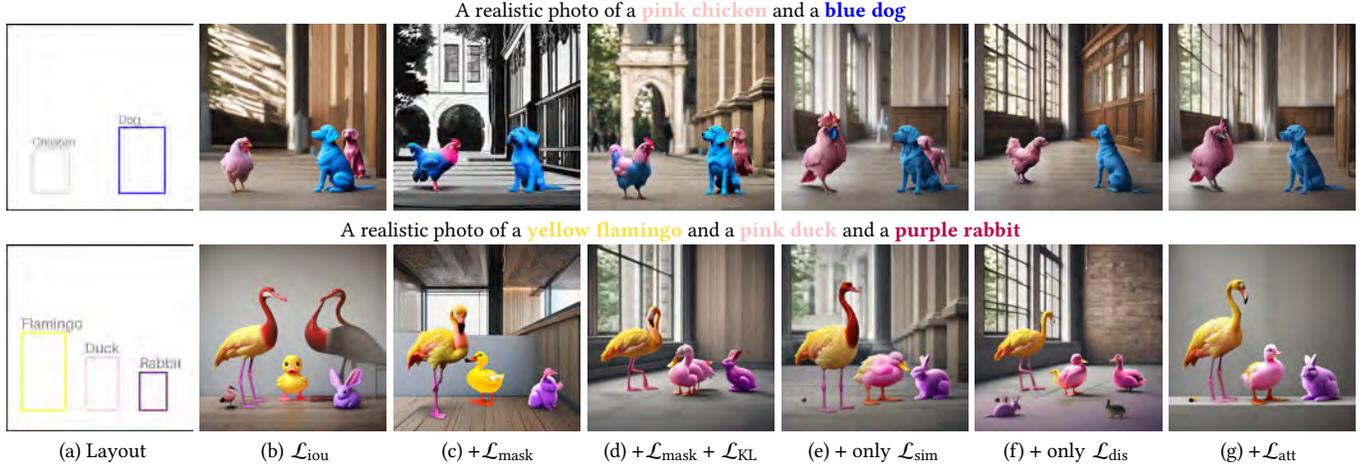


Fig. 6. Qualitative ablation on the impact of our various loss terms in Eq. (11). From L2R, columns are: (a) layout prompt, (b) output of \mathcal{L}_{iou} only, (c) effect of masked latent regularization ($\mathcal{L}_{iou} + \mathcal{L}_{mask}$), and (d) together with KL regularization ($\mathcal{L}_{iou} + \mathcal{L}_{mask} + \mathcal{L}_{KL}$). Next, we present the impact of adding attribute losses to (d): (e) similarity loss ($+\mathcal{L}_{sim}$), (f) dissimilarity loss ($+\mathcal{L}_{dis}$), and (g) full attribute loss ($+\mathcal{L}_{att}$) consisting of all loss terms, and corresponding to our final approach, MALeR. All images are generated using seed 0.

Table 4. Ablation of individual loss components on DrawBench.

	Losses				Spatial Acc.	Color Acc.	Counting		
	\mathcal{L}_{iou}	\mathcal{L}_{mask}	\mathcal{L}_{KL}	\mathcal{L}_{att}			P	R	F1
1	✓	-	-	-	0.69	0.44	0.74	0.95	0.83
2	✓	✓	-	-	0.63	0.39	0.78	0.85	0.81
3	✓	-	✓	-	0.74	0.44	0.75	0.96	0.84
4	✓	✓	✓	-	0.81	0.47	0.81	0.96	0.88
5	✓	✓	✓	✓	0.81	0.61	0.81	0.96	0.88

tasks (row 3), highlighting its role in stabilizing latent representations. Interestingly, when both loss terms are applied together, the method achieves highest performance by simultaneously mitigating background leakage and stabilizing the latent optimization process (row 4). Finally, including the attribute loss (\mathcal{L}_{att}) results in a significant improvement to color accuracy (row 5).

We show the qualitative impact of masked regularization, KL alignment loss, and components of our attribute loss, in Fig. 6. Only using the masked latent regularization term effectively prevents background semantic leakage, but results in reduced visual fidelity, e.g. a black-and-white background (row 1) or a slightly cartoonish duck (row 2). When combined with the KL alignment term, image quality improves while background semantic leakage continues to be suppressed. However, attribute leakage is observed, e.g. the blended pink and blue chicken (row 1) and out-of-shape duck (row 2). Among the subject-attribute association losses, only including one of the similarity or dissimilarity loss (columns e, f) fails to resolve attribute leakage. In fact, we also see some subject leakage with the chicken having a dog’s head (col f, row 1). In contrast, our final formulation with all loss terms (masked regularization, KL alignment, and both similarity and dissimilarity subject-attribute losses), results in correct subject identities with accurate attributes bound to each entity (col g). These qualitative observations support the quantitative results presented in Tab. 4.

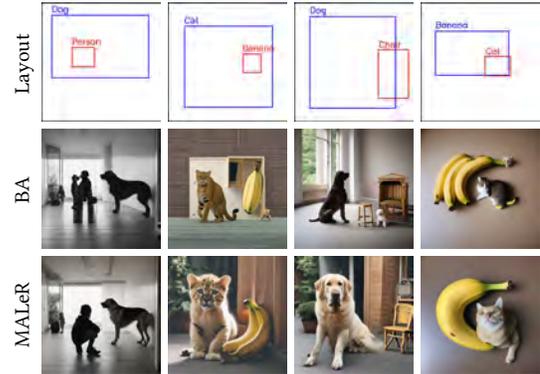


Fig. 7. **Limitations.** MALeR struggles to generate images when subjects are completely overlapping or have atypical aspect ratios. However, generations for partial overlaps are acceptable. L2R prompts from HRS: (a) a small person and a big dog; (b) a small banana and a big cat; (c) a big dog and a small chair; and (d) a large banana and a small cat.

4.4 Limitations

While MALeR is effective in compositional scene generation using layout guidance, it has certain limitations. First, its performance is inherently constrained by the generative capabilities of SDXL, which may result in suboptimal images where SDXL faces challenges. Second, while our method performs reasonably well with partially overlapping bounding boxes, it may produce images that do not adhere to the layout in cases involving fully overlapping bounding-boxes (Fig. 7). Atypical aspect ratios due to automatically generated bounding box layouts (e.g. on HRS) are challenging, but unlikely with real user interactions. In column 1, our method fails to generate a small person in front of a large dog; while in column 2, the banana appears larger than the specified bounding box. However, in columns 3 and 4, our method performs reasonably well with partially overlapping boxes. Future advancements in dealing with overlapping boxes may improve robustness in such cases.

5 Conclusion

We presented MALeR, a training-free, layout-guided text-to-image approach that enables users to generate compositional scenes involving multiple subjects and multiple attributes. We pointed out three primary challenges in existing approaches: (i) background semantic leakage, (ii) out-of-distribution generations, and (iii) inaccurate subject-attribute binding in compositional scenes. MALeR addressed these challenges through (i) masked latent regularization, (ii) in-distribution latent alignment, and (iii) a subject-attribute association loss. Quantitative comparison showed improved performance over existing approaches while image visualizations showed the ability of MALeR to generate accurate, controllable, and compositional images, with enhanced stability and consistent attribute binding. We confirmed that MALeR is perceived to generate better images that are more robust to random seeds through user studies.

Acknowledgments

This project was supported by funding from an Adobe Research gift. MT also thanks SERB SRG/2023/002544 grant for compute support. We thank all volunteers who participated in our user study.

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 *International Conference on Computer Vision (ICCV)*.
- 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. In *International Conference on Computer Vision (ICCV)*.
- Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324* (2022).
- Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Universal Guidance for Diffusion Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *International Conference on Machine Learning (ICML)*.
- Barak Battash, Amit Rozner, Lior Wolf, and Ofir Lindenbaum. 2024. Obtaining Favorable Layouts for Multiple Object Generation. *arXiv preprint arXiv:2405.00791* (2024).
- 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. 2024. Training-Free Layout Control with Cross-Attention Guidance. In *Winter Conference on Applications of Computer Vision (WACV)*.
- Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. 2024. Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation. In *European Conference on Computer Vision (ECCV)*.
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.
- Yuki Endo. 2024. Masked-Attention Diffusion Guidance for Spatially Controlling Text-to-Image Generation. *The Visual Computer* 40, 9 (2024), 6033–6045.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *International Conference on Machine Learning (ICML)*.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *International Conference on Learning Representations (ICLR)*.
- Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. 2024. Ranni: Taming Text-to-Image Diffusion for Accurate Instruction Following. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. MDTV2: Masked Diffusion Transformer is a Strong Image Synthesizer. In *International Conference on Computer Vision (ICCV)*.
- Yuchao Gu, Yipin Zhou, Yunfan Ye, Yixin Nie, Licheng Yu, Pingchuan Ma, Kevin Qinghong Lin, and Mike Zheng Shou. 2025. ROIctrl: Boosting Instance Control for Visual Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. 2024. InitNo: Boosting Text-to-Image Diffusion Models via Initial Noise Optimization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing Prompts for Text-to-Image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- 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).
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*.
- Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and Controllable Image Synthesis with Composable Conditions. In *International Conference on Machine Learning (ICML)*.
- Eric Hanchen Jiang, Yasi Zhang, Zhi Zhang, Yixin Wan, Andrew Lizarraaga, Shufan Li, and Ying Nian Wu. 2024. Unlocking the Potential of Text-to-Image Diffusion with PAC-Bayesian Theory. *arXiv preprint arXiv:2411.17472* (2024).
- Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 1 (1951), 79–86.
- Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. 2023a. Divide & Bind Your Attention for Improved Generative Semantic Nursing. In *British Machine Vision Conference (BMVC)*.
- 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, and Kwan-Yee K Wong. 2024. Place: Adaptive Layout-Semantic Fusion for Semantic Image Synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. 2024. Dynamic Prompt Optimizing for Text-to-Image Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning (ICML)*.
- Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *International Conference on Machine Learning (ICML)*.
- Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. 2024. Compositional Text-to-Image Generation with Dense Blob Representations. In *International Conference on Machine Learning (ICML)*.
- William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Quynh Phung, Songwei Ge, and Jia-Bin Huang. 2024. Grounded Text-to-Image Synthesis with Attention Refocusing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations (ICLR)*.
- 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. In *ACM Multimedia (MM)*.

- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*.
- 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. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.
- 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. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. 2024. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. In *SIGGRAPH Asia Conference Papers*.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021a. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*.
- Aravindan Sundaram, Ujjayan Pal, Abhimanyu Chauhan, Aishwarya Agarwal, and Srikrishna Karanam. 2024. CoCoNO: Attention Contrast-and-Complete for Initial Noise Optimization in Text-to-Image Synthesis. *arXiv preprint arXiv:2411.16783* (2024).
- Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In *Association of Computational Linguistics (ACL)*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. 2024. InstanceDiffusion: Instance-Level Control for Image Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Sam Witteveen and Martin Andrews. 2022. Investigating Prompt Engineering in Diffusion Models. *arXiv preprint arXiv:2211.15462* (2022).
- Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. 2023. Harnessing the Spatial-Temporal Attention of Diffusion Models for High-Fidelity Text-to-Image Synthesis. In *International Conference on Computer Vision (ICCV)*.
- Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. 2024. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240* (2024).
- Jiayu Xiao, Henglei Lv, Liang Li, Shuhui Wang, and Qingming Huang. 2024. R&B: Region and Boundary Aware Zero-Shot Grounded Text-to-Image Generation. In *International Conference on Learning Representations (ICLR)*.
- Jinhang 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 *International Conference on Computer Vision (ICCV)*.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. 2023. ReCo: Region-Controlled Text-to-Image Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *International Conference on Computer Vision (ICCV)*.
- Peiang Zhao, Han Li, Ruiyang Jin, and S Kevin Zhou. 2023. LoCo: Locally Constrained Training-Free Layout-to-Image Synthesis. *arXiv preprint arXiv:2311.12342* (2023).
- Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-Image Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. 2024. Fast Training of Diffusion Models with Masked Transformers. *Transactions on Machine Learning Research (TMLR)* (2024).
- Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. 2024. Migc: Multi-instance generation controller for text-to-image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

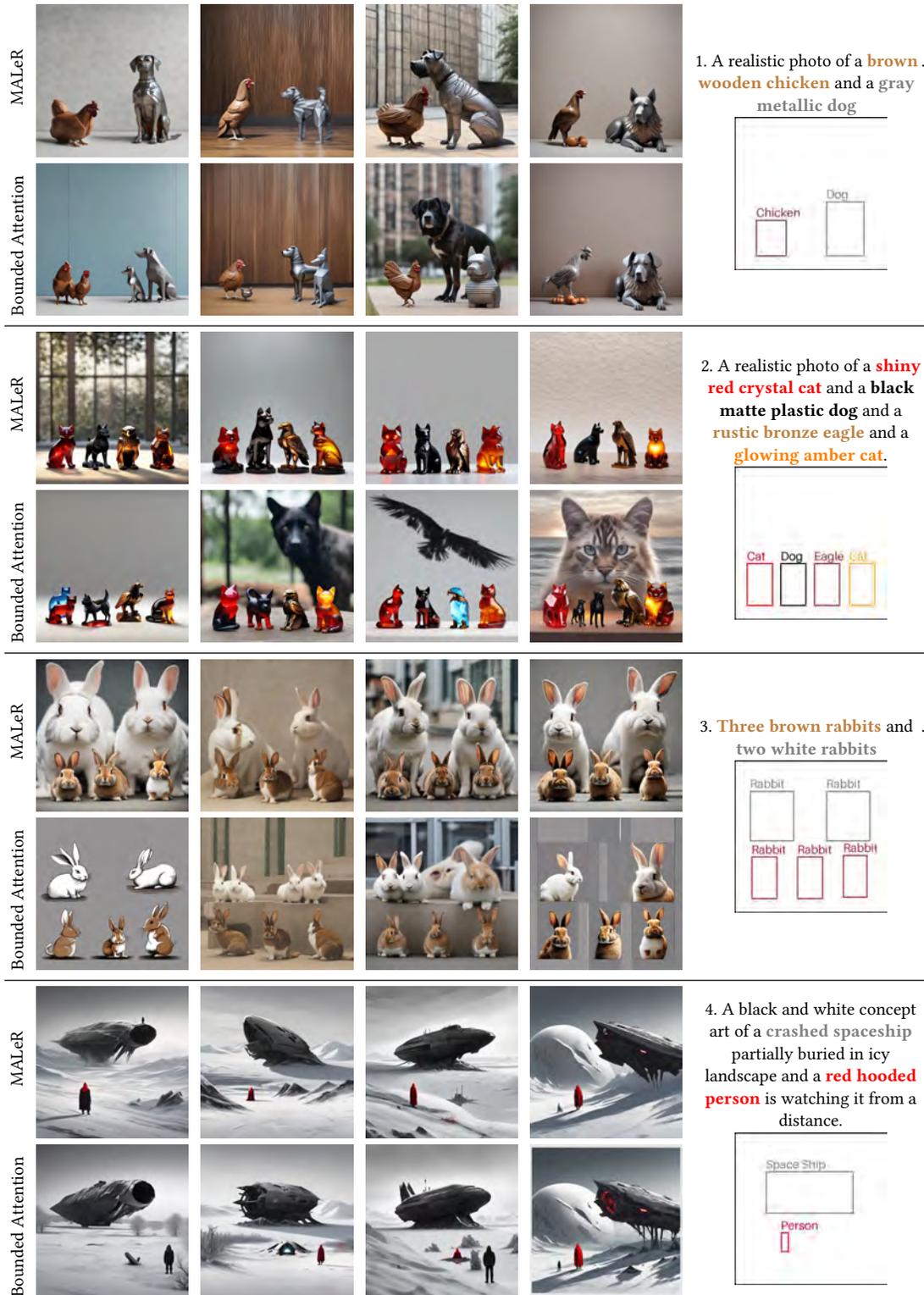
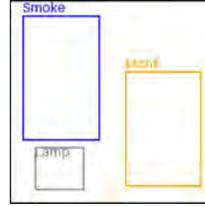


Fig. 8. **Effect of random seeds.** We present 4 images for the same input text and layout prompt only differing based on the initial random seed. Within a column, both Bounded Attention and MALeR use the same seed. We see that MALeR generates more accurate images while Bounded Attention suffers many problems mentioned in the paper. Background semantic leakage is prominently seen in 3 of 4 seeds for prompt 2, unnatural out-of-distribution artifacts for all four generations of prompt 3 (rabbits), and erroneous attributes are seen across multiple generations.

A **monk wearing orange robes** and his **dog** together on cliff side and **pink cherry blossoms** in japanese sumi-e ink style painting



An ancient japanese calligraphy sketch of a **blue mystic smoke** coming out of a small metallic **lamp** and a **monk wearing orange robe** sitting nearby



A concept art of a icy landscape with a **red wizard** standing far away firing his wand in air for a **black colored magic thunderbolt**



A concept art of a bright desert with **two magical portals** in air. One portal showing modern city with skyscrapers and another portal showing **black and white apocalyptic city** with broken buildings



A concept art of a epic duel with a far away **green robed wizard** firing **magic thunderbolt** in air towards a **wizard** protecting himself with his magical shield in an surreal icy windy landscape



A concept art of a black and white place with a **green robe wizard** far away firing his wand in air for a **blue thunderbolt** with a sheer force



A realistic photo of a **purple rabbit** and a **pink chicken** and a **yellow cat**



A realistic photo of a **purple cat** and a **pink cat** and a **black cat**

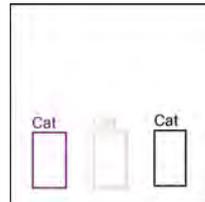


Fig. 9. **Highlighting robustness of MAlER.** We present 8 qualitative results on diverse and challenging prompts with complex layout constraints to showcase the ability of MAlER to synthesize high-quality, coherent, and spatially accurate images.