SPADIFF: SPATIAL CONDITIONING ADAPTER FOR OPEN-VOCABULARY SEGMENTATION-TO-IMAGE

Anonymous authors
Paper under double-blind review



Figure 1: **SpaDiff brings precise open-vocabulary spatial control to FLUX** without training the base model. SpaDiff generates photorealistic images from a segmentation mask along with corresponding regional prompts, and a global prompt (*left*), and naturally composes with other adapters, like Canny and Depth (*right*).

ABSTRACT

When creating visual designs, a highly desired feature is precise spatial control over image generation. Segmentation-mask-to-image (S2I) generation provides such control by binding a textual description, or regional prompt, to each specific image region present in the segmentation mask during generation. While open-vocabulary S2I methods allow arbitrary prompts, they fine-tune the entire base model, so other extensions trained on the original weights no longer combine with them reliably. Adapters, lightweight modules trained on top of a frozen base model, preserve this composability, but existing S2I adapters are restricted to closed vocabularies and built on older base models, transferring poorly to modern Multi-Modal Diffusion Transformers (MM-DiTs). We present **SpaDiff**, an open-vocabulary S2I adapter for the MM-DiT FLUX that keeps the base model frozen. SpaDiff combines two mechanisms: (1) an adaptive attention masking strategy that binds regional prompts to their corresponding image region by suppressing image and text tokens interactions softly with a learned per-timestep, per-layer, and per-head strength to avoid disrupting the frozen base model; and (2) a mechanism for adhering to exact mask boundaries, explored in two variants, along with a lightweight feature gating mechanism. SpaDiff achieves state-of-the-art performance among adapter-based methods on both the closed-vocabulary COCO-Stuff and open-vocabulary SACap-Eval benchmarks, while approaching the performance of methods that fine-tune the full base model.

1 INTRODUCTION

The automated generation of photorealistic imagery opens up new workflows for expert designers such as architects or the general public. Most of these workflows are powered by Diffusion Models (DMs) (Song et al., 2020; Ho et al., 2020), and in particular by text-to-image (T2I) generation (Dhariwal & Nichol, 2021; Rombach et al., 2022; Esser et al., 2024; Labs). Text alone, however, gives no precise control over the exact positions of elements and their visual details, which is essential for visual design tasks such as architectural rendering. This gave rise to methods that condition the diffusion process on reference images (Ye et al., 2023) or spatial layouts (Zhang et al., 2023; Mou et al., 2024; Zhao et al., 2023; Li et al., 2024; Tan et al., 2025). Particularly useful are segmentation-mask-to-image (S2I) methods (Lv et al., 2024; Xue et al., 2023; Li et al., 2025; Zhang

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

et al., 2025a), which take a segmentation mask with a text description per entity and thus specify both the position and the semantic content of each element. For even finer control, experts additionally require multi-conditioning (Zhang et al., 2023) on further spatial cues, such as canny edges or depth maps alongside the segmentation map.

Three properties determine the usability of an S2I model. First, the vocabulary should be open rather than closed: instead of a fixed set of class labels available at inference time, the user should be free to describe any entity in arbitrary text. Second, the method should be modular. Baked-in methods fine-tune the backbone to integrate the conditional signal, which pushes every co-loaded module out of the input distribution it was trained and validated on, whereas adapters keep the base model frozen, train only a small set of parameters, and can therefore be combined trivially. Third, it should build on a state-of-the-art diffusion model such as FLUX.1 (Labs), which offers high image quality and prompt adherence. No existing method satisfies all three. S2I adapters for older UNet-based models (Rombach et al., 2022), such as ControlNet (Zhang et al., 2023), are closed-vocabulary and yield suboptimal results on newer MM-DiTs (Peebles & Xie, 2023; Esser et al., 2024; Tan et al., 2025). Baked-in methods such as PLACE (Lv et al., 2024) and FreestyleNet (Xue et al., 2023) are open-vocabulary but share those older UNet backbones. Recent open-vocabulary methods built on FLUX.1, such as Seg2Any (Li et al., 2025) and EliGen (Zhang et al., 2025a), remain baked-in.

We propose SpaDiff, an adapter for the MM-DiT architecture that enables open-vocabulary S2I without fine-tuning the base model. SpaDiff has two main mechanisms: (1) an adaptive attention mask that binds each entity's textual description to its spatial position in the segmentation mask, while maintaining an adapter design; and (2) a conditional integration method that guides the base model to respect the contours of the segmentation map, while keeping the base model frozen. Unlike Seg2Any's hard attention masking, our adaptive masking lets the model learn how strongly to block token interactions depending on the denoising timestep, transformer layer, and attention head. For conditional integration we explore two variants (Fig. 2). In the unified variant, the conditional image becomes a new modality within MM-DiT, with a separate pathway symmetrical to those of the text and noisy image, and its tokens are concatenated with the text and image tokens for the attention operation. In the decoupled variant, the conditional tokens instead pass through a separate attention with only the image tokens, whose output is added to the image branch; to stabilize its training we introduce a ZeroScale layer for feature gating, inspired by ControlNet's ZeroConv. We contribute as follows:

- We introduce SpaDiff, an open-vocabulary adapter-style S2I model built on the FLUX architecture that does not require fine-tuning the backbone.
- We propose a learnable adaptive masking mechanism, two variants of conditional image integration (unified and decoupled), and a ZeroScale layer for parameter-efficient feature gating.
- SpaDiff achieves SOTA performance among adapters on both closed-vocabulary and open-vocabulary S2I benchmarks, surpassed only by its baked-in, rigid counterpart, Seg2Any.

2 Related Works

Diffusion Transformers Modern text-to-image (T2I) generation has shifted both in training recipe and architecture. Latent Diffusion Models (LDMs) (Rombach et al., 2022), which denoise in the latent space of a pre-trained VAE (Kingma & Welling, 2013), originally used convolutional UNets trained with denoising score matching (Song et al., 2020; Ho et al., 2020). Diffusion Transformers (DiTs) (Peebles & Xie, 2023) replaced the UNet with a more scalable transformer operating on patchified latents, while flow matching (Lipman et al., 2022) and rectified flow (Liu et al., 2022) replaced the score/noise-matching objective. Multi-Modal Diffusion Transformers (MM-DiTs) (Esser et al., 2024) extend DiTs by processing text and image tokens through separate parameter streams that interact via joint self-attention. FLUX.1 (Labs) combines rectified-flow training with an MM-DiT architecture and serves as the base model for SpaDiff.

Adapters are lightweight trainable modules that add controllability to pre-trained T2I models while keeping the base model frozen (Zhang et al., 2023; Mou et al., 2024; Ye et al., 2023). They enable spatial control through conditions such as edges, depth, and segmentation maps (Zhang et al.,

2023; Mou et al., 2024), as well as appearance control through reference images (Ye et al., 2023). Early adapters were designed for UNet-based DMs (Zhang et al., 2023; Mou et al., 2024; Ye et al., 2023; Zhao et al., 2023), whose integration mechanisms do not directly transfer to MM-DiTs (Esser et al., 2024). OminiControl (Tan et al., 2025) addresses this architectural shift by adding a new parameter stream for the conditional image identical to the existing text and image streams. SpaDiff investigates this unified strategy alongside a decoupled attention variant inspired by IP-Adapter (Ye et al., 2023), while targeting open-vocabulary S2I, a setting existing adapters do not support.

Segmentation-mask-to-image generation Current adapter-based S2I methods, such as ControlNet's (Zhang et al., 2023) segmentation variant and T2I-Adapter (Mou et al., 2024), treat the segmentation mask as a generic spatial conditioning signal, encoding class identity through visual appearance (e.g., color codes) rather than text. This restricts them to a closed label set, with no mechanism for binding arbitrary descriptions to regions. In contrast, open-vocabulary S2I methods (Lv et al., 2024; Xue et al., 2023; Zhang et al., 2025a) introduce attention masking that binds text tokens to specific spatial regions. However, because attention operates on coarse latent patches, they cannot capture fine object boundaries. Seg2Any (Li et al., 2025) addresses this trade-off by combining attention masking with a contour map of the segmentation mask, supplied as an additional conditioning image. Unlike Seg2Any, which fine-tunes the full backbone, SpaDiff achieves the same open-vocabulary S2I capability through a frozen-backbone adapter design.

3 METHOD

The goal of open-vocabulary S2I is to generate an image conditioned on a global caption p_0 and N regional captions p_i, each assigned to a region s_i in a segmentation map $S \in \mathbb{R}^{H\times W\times 3}$: $[p_0, (p_1, s_1), \ldots, (p_N, s_N)]$. Here, each $s_i \subseteq 1, \ldots, H \times 1, \ldots, W$ denotes one distinctly colored region in S, and p_i is a free-form textual description of the corresponding entity.

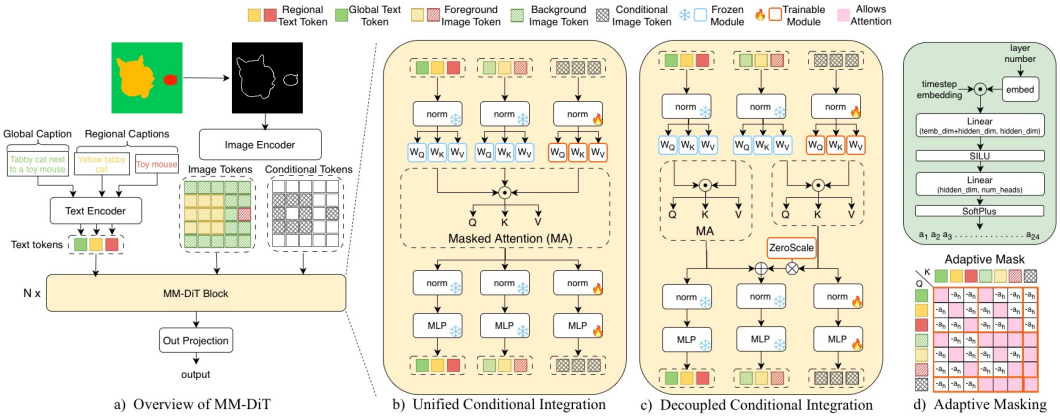

Figure 2: **Overview of SpaDiff**: a) The global and regional captions get encoded and concatenated into text tokens. Contours are extracted from the segmentation map and get encoded into conditional tokens, while noisy image tokens are mapped to their corresponding regions. b) A conditional stream gets added to the MM-DiT backbone and the conditional tokens are processed with the image and text tokens through the same Joint-Attention operation. c) Different from the Unified approach, the conditional tokens get processed through a separate Joint-Attention operation with just the image tokens and the output is feature gated by a ZeroScale layer and added to the image stream. d) The values within the adaptive attention mask that is applied to every attention operation are computed as a learned function of the timestep embedding and layer number withing the MM-DiT architecture. There is one value per attention head.

3.1 UNIFIED AND DECOUPLED CONDITIONAL INTEGRATION

We follow Seg2Any for processing a segmentation map S. We start by extracting the contours to generate a corresponding contour map, which is then encoded using the same VAE image encoder as for the noisy image, to obtain a set of patch-level tokens (Fig. 2(a)). The tokens that do not contain a contour are dropped for computational efficiency. To integrate the conditional tokens, we add a

third stream $\mathcal{C}$ symmetric to the existing text $\mathcal{T}$ and image $\mathcal{I}$ streams. Each stream $m \in \{\mathcal{I}, \mathcal{T}, \mathcal{C}\}$ has its own projection matrices W_Q^m, W_K^m, W_V^m that produce per-modality queries, keys, and values (Q_m, K_m, V_m). These are concatenated along the sequence dimension over a chosen subset of modalities to form Q, K and V, and subsequently passed to a masked scaled dot-product attention:

$$A(Q, K, V, M) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \log(M)\right) V, \tag{1}$$

where M is an attention mask. We instantiate the attention operation in two ways.

Unified All three streams are concatenated and processed by a single attention operation, yielding a single output $A(Q, K, V, M)$ that updates all tokens jointly (Fig. 2(b)).

Decoupled Two attention operations run in parallel: $A_{\mathcal{TI}} = A(Q_{\mathcal{TI}}, K_{\mathcal{TI}}, V_{\mathcal{TI}}, M_{\mathcal{TI}})$ over the text and image streams, and $A_{\mathcal{IC}} = A(Q_{\mathcal{IC}}, K_{\mathcal{IC}}, V_{\mathcal{IC}}, \mathbf{1})$ over the image and conditional streams (Fig. 2(c)). The second operation uses an all-ones mask (i.e., no blocking), since image and conditional tokens are unrestricted in their interaction. The image portions of $A_{\mathcal{TI}}$ and $A_{\mathcal{IC}}$ are summed before the residual so that image tokens receive information from both pathways while text and conditional tokens are updated only through their respective attention operations. The image tokens become:

$$A^{\text{img}} = A^{\text{img}}_{\mathcal{TI}} + c\,\boldsymbol{\gamma} A^{\text{img}}_{\mathcal{IC}}, \tag{2}$$

where c is the conditioning strength and $\boldsymbol{\gamma}$ is the ZeroScale applied per-feature (element-wise) across all tokens (see Section 3.2 for details on ZeroScale).

Strength Modulation The conditioning strength c allows for stricter or looser adherence to the given layout. In the unified case, we follow Seg2Any's approach to add a bias matrix into the attention operation, by assigning c as the value between conditional and image tokens in the attention mask M. In the decoupled case we directly adjust the influence of the conditional tokens when adding them to the image path, as seen in Eq. (2).

Adapter Design Crucially, the modules that process the conditional tokens are the only ones being finetuned, thus keeping the base FLUX model frozen, which allows for stacking multiple adapters together to achieve multi-conditioning. Moreover, we follow OminiControl's (Tan et al., 2025) parameter reuse strategy for the symmetrical modules and use a small LoRa (Hu et al., 2022) adapter applied to the image stream that only gets enabled when processing the conditional tokens.

3.2 ZeroScale

The rigid addition in the decoupled setting introduces graininess in the generated images (Fig. 4). We therefore gate the conditional contribution with a ZeroScale layer: a learnable per-feature scaling vector $\boldsymbol{\gamma} \in \mathbb{R}^d$ (Eq. (2)), initialized to zero so the conditional pathway starts inactive and is introduced gradually over training. ZeroScale is inspired by ControlNet's zero-convolution (Zhang et al., 2023), but replaces its $d \times d$ mixing matrix with a per-feature gate, which is sufficient to modulate the pathway: across FLUX's 57 layers this costs ≈ 175K parameters instead of ≈ 537M. We train the decoupled stream first and only then fine-tune it jointly with ZeroScale for fewer steps, which lets the gate open further because the features it admits are already meaningful (Section 4.2.1).

3.3 Adaptive Masking

Open-vocabulary S2I binds each regional prompt p_i to its corresponding image region by masking text–image attention according to the segmentation map (Eq. (1) and Fig. 2). Seg2Any (Li et al., 2025) uses a hard mask, assigning $-\infty$ to disallowed interactions; while effective after full backbone fine-tuning, applying this directly to frozen FLUX disrupts its learned attention distributions and performs poorly (Section 4.2.2). We therefore replace hard masking with an adaptive soft mask: each blocked interaction receives a learned penalty $-a_{t,\ell,h}$, $a > 0$, where the suppression strength depends on denoising timestep t, transformer layer ℓ, and attention head h. This reflects that layout emerges differently over denoising timestep, across FLUX layers (Zhou et al., 2025), and across attention heads. A single value $a_{t,\ell,h}$, learned by a lightweight module (Fig. 2(d)) is shared across

all disallowed token pairs at a given (t, ℓ, h). Thus, rather than forcing interactions completely off, SpaDiff learns when, where, and how strongly to suppress them while leaving the backbone frozen. For decoupled conditional integration, masking is applied only to text–image attention $A_{\mathcal{TI}}$; image–condition attention $A_{\mathcal{IC}}$ remains unrestricted.

4 EXPERIMENTS

Datasets Experiments are conducted on the open-set segmentation dataset SACap-1M (Li et al., 2025), as well as the closed-set segmentation dataset COCO-Stuff (Caesar et al., 2018). For COCO-Stuff, regional captions are assigned as the semantic class names of each segment and no global caption is used unless indicated.

Metrics We use the mean intersection-over-union (MIoU) between predicted and input segmentation map to measure the degree to which the model respects the input segmentation. On COCO-Stuff, we use DeepLabV3 (Chen et al., 2017) to compute the semantic segmentation of the generated image; on SACap-Eval, we use SAM2 (Ravi et al., 2024) and report class-agnostic MIoU. For global image quality, we report FID (Heusel et al., 2017), which measures the distance between the test and generated image distributions, and MAN-IQA (Yang et al., 2022), which assesses perceptual quality. When a global caption is available, we measure CLIP Score that shows the alignment between the generated image and the global text, PickScore (Kirstain et al., 2023), and Image Reward (IR) (Xu et al., 2023), which both predict human preference. On SACap-Eval, we further evaluate regional quality using Spatial, Color, Shape and Texture Scores (Zhang et al., 2025b), which measure the alignment between each image region and its corresponding regional caption, using Qwen2-VL-7B (Wang et al., 2024) in a Visual Question Answering (VQA) manner.

Implementation details All experiments are conducted using FLUX.1-dev (Labs). Our method uses LoRA (Hu et al., 2022) with rank 64 for the conditional integration, which is only active on the stream processing the conditional tokens.

Training All conditional integration methods are trained for 20k steps, while the ZeroScale is trained for 5k steps in the best setting. When combining the conditional integration with adaptive masking, the adaptive mask is pre-trained for 20k steps and then finetuned for 20k more steps together with the conditional integration. We use an effective batch size of 12 and AdamW optimizer with a fixed learning rate of e^{-4}, except for training the ZeroScale, where the learning rate is ablated. All experiments are conducted on NVIDIA A40, L40 or A100 GPUs.

Baselines We compare against a broad set of baselines. Among Stable Diffusion 1.5-based methods, ControlNet (Zhang et al., 2023) and ControlNet++ (Li et al., 2024) are closed-set adapters, while PLACE (Lv et al., 2024) and FreestyleNet (Xue et al., 2023) are open-set methods that fine-tune the full backbone. Among more recent FLUX-based methods, EliGen (Zhang et al., 2025a) and Seg2Any (Li et al., 2025) likewise fine-tune the full backbone for open-set generation.

4.1 MAIN RESULTS

Table 1 and Fig. 3(a) show quantitative and qualitative results on closed-vocabulary COCO-Stuff. UNet-based adapters struggle to respect the input segmentation and occasionally produce unrealistic images, while the remaining methods achieve competitive scores. Visually, Seg2Any and SpaDiff stand out due to their FLUX backbone, producing noticeably higher quality. Their visual quality is comparable, but Seg2Any achieves the highest MIoU (45.67), with SpaDiff a close second (44.13) at a better FID.

Table 2 and Fig. 3(b) present results on the open-vocabulary SACap-Eval benchmark. UNet-based methods score lowest across almost all metrics, struggling with realism and precise entity placement, particularly for small objects. EliGen by design supports only loose positioning and thus obtains a low MIoU, though it respects the input regional prompts well; visually, its images are less realistic and more cartoonish. Seg2Any and SpaDiff far outperform the other methods, reaching MIoU of 95.03 and 85.41 respectively, closely respecting object positioning and shapes. Note that the gap between the two is also influenced by the benchmark's 1024x1024 resolution, whereas SpaDiff was

trained at 512x512. In terms of global quality, the FLUX backbone is essential for producing visually pleasing images and processing the regional text prompts. This is also reflected in the regional quality metrics, where SpaDiff slightly outperforms Seg2Any, while visually the two produce very similar results that surpass all other baselines.

Table 1: **Results on COCO-Stuff**, where green and blue represent the best and second best methods, respectively. SpaDiff is second-best in terms of MIoU and better in terms of FID than Seg2Any.

Method	MIoU↑	Global quality	
		FID↓	MAN-IQA↑
ControlNet++	19.9	30.13	55.8
FreestyleNet	43.07	17.65	54.9
PLACE	42.85	17.22	51.1
Seg2Any	45.67	20.88	56.9
SpaDiff	44.13	20.02	54.6

Table 2: **Results on SACap1M.** SpaDiff achieves high MIoU close to Seg2Any, with strong regional prompt adherence and high global quality due to the FLUX backbone.

Method	MIoU↑	Global quality					Regional quality			
		FID↓	MAN-IQA↑	CLIP↑	Pick↑	IR↑	Spatial↑	Color↑	Shape↑	Texture↑
FreestyleNet	74.54	25.23	58.11	24.80	19.94	-1.01	54.99	37.05	39.61	40.79
PLACE	71.79	23.38	60.11	25.49	20.09	-0.84	57.52	41.36	43.89	44.79
EliGen	52.86	23.14	50.85	27.95	22.30	0.77	79.84	70.06	70.64	72.07
Seg2Any	95.03	17.88	50.69	27.72	21.66	0.40	87.17	78.30	78.51	79.80
SpaDiff	85.41	15.71	48.75	27.84	22.16	0.63	87.30	79.61	79.68	80.90

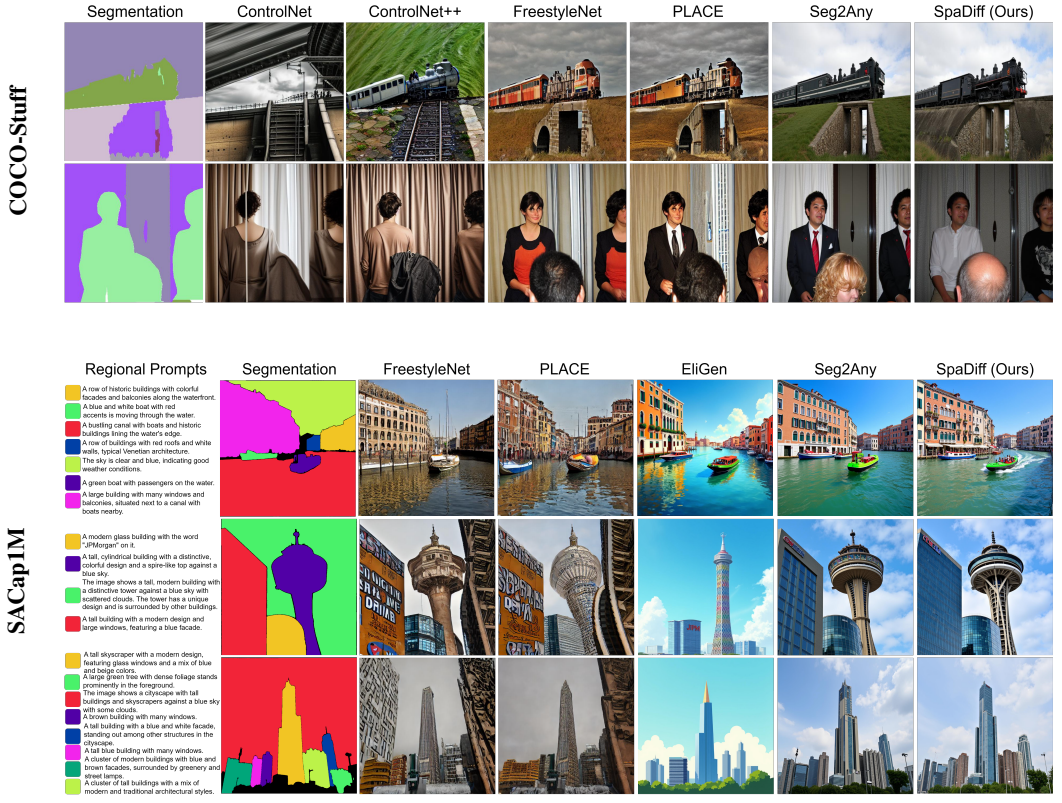

Figure 3: **Qualitative examples on COCO-Stuff and SACap1M.** SpaDiff closely respects the input segmentation, outperforming UNet-based adapter and baked-in methods on COCO-Stuff and UNet-based methods and EliGen on SACap1M, while matching Seg2Any in visual quality.

4.2 Ablation Studies

4.2.1 Conditional Integration

ZeroScale ablations We ablate the training of the ZeroScale layer along two control knobs: the training methodology and the ZeroScale learning rate. *Together* means ZeroScale and the decoupled conditioning are trained jointly from scratch; *phases* means the decoupled conditioning is trained first and used as initialization for ZeroScale training. As shown in Table 3, training together removes the graininess in generated images (reflected in MAN-IQA) but hurts MIoU and FID. The cause is visible in Fig. 4: the gate barely opens, with ZeroScale parameters remaining near zero throughout training, suggesting the gate only opens once the decoupled stream produces meaningful features. Training in phases fixes both scores, but properly opening the gate additionally requires a higher ZeroScale learning rate. While Fig. 4 suggests that a higher learning rate simply opens the gate more, Table 3 shows this is not always desirable and that a fine balance must be maintained, further confirmed by a learning rate scheduler experiment in which the gate starts closing as the learning rate decreases. The final decoupled conditioning mode uses ZeroScale trained in phases with a x100 learning rate.

Table 3: **ZeroScale ablation on COCO-Stuff.** The decoupled conditioning branch is always trained with 10^{-4}, while the ZeroScale learning rate uses the same base value. A learning-rate scheduler gives the best FID, while a constant $\times 100$ learning rate improves MIoU by more than one point.

Training	Learning rate	Scheduler	MIoU↑	Global quality	
				FID↓	MAN-IQA↑
			42.96	21.43	50.4
together	$\times 1$		37.58	27.14	54.3
phases	$\times 1$		41.06	26.32	54.3
phases	$\times 100$		43.82	21.74	54.2
phases	$\times 250$		41.41	21.46	56.4
phases	$\times 500$		41.85	22.58	53.5
phases	$\times 500$	linear	42.61	21.15	53.2



Figure 4: The difference between a grainy image obtained in the decoupled setting and a clean image obtained when introducing ZeroScale.

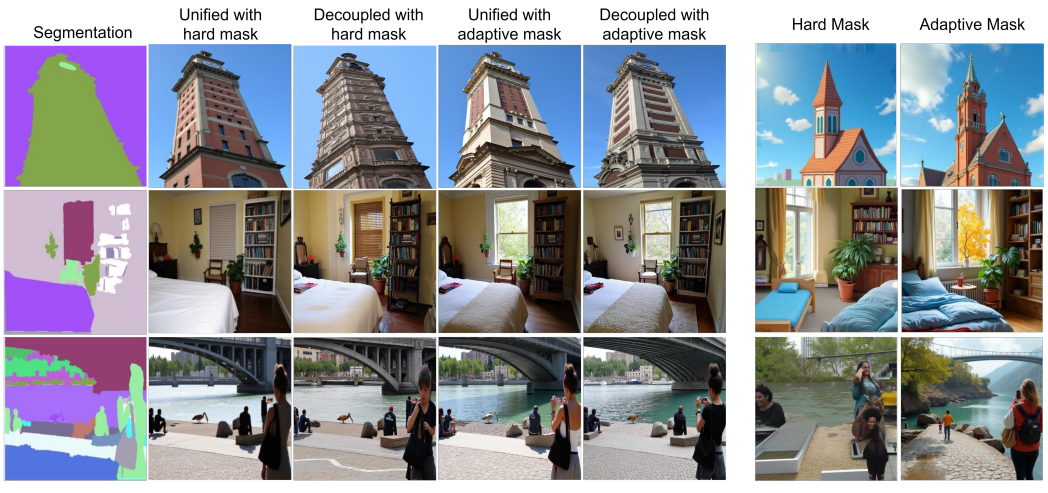

Figure 5: **Effect of the type of integration (column 2 - 5) and masking (column 6 and 7).** Adaptive masking performs best, with unified and decoupled integration yielding similar results.

Conditional Integration and Attention Masking Table 4 ablates the combination of SpaDiff's two key mechanisms: conditional integration (decoupled vs. unified) and attention masking (adaptive vs. hard). The unified setting outperforms the decoupled in MIoU, while visual quality remains

comparable (Fig. 5). Adaptive masking consistently improves FID and gives a slight MIoU boost in the unified setting. Without conditional integration, the adaptive mask still produces high-quality images and achieves loose positioning of the dominant entities, whereas the hard mask produces un-cohesive ones (Fig. 5, Table 5). The best configuration, used by SpaDiff, combines unified integration with adaptive masking.

Table 4: **Effect of conditional integration and attention masking combinations.** Unified always outperforms decoupled in terms of MIoU, while adaptive masking substantially improves FID.

Conditional Integration		Attention Masking		MIoU↑	Global quality	
unified	decoupled	hard	adaptive		FID↓	MAN-IQA↑
✓		✓		43.9	23.46	54.4
	✓	✓		43.82	21.74	54.2
✓			✓	44.13	20.02	54.6
	✓		✓	43.67	20.09	54.3

4.2.2 Adaptive Masking ablations

We ablate the three factors influencing the attention masking bias: timestep, layer index and attention head. Table 5 shows that the full adaptive masking scheme using all three performs best. The combination of timestep and layer index is the most crucial, with the per-head values providing significant improvements in MIoU and FID. These ablations also let us study each factor in isolation (Fig. 6). The higher mask values at early denoising timesteps affirm that diffusion models decide the layout of the generated image early on. In the no per-head setting, the mask strength reveals that some final and middle layers of the FLUX architecture contribute most to the layout, while the mean strength per attention head shows that some heads play a much larger role in spatial layout than others. Together, these findings show that layout control is concentrated in specific timesteps, layers and attention heads, justifying an adaptive mask that learns where to suppress rather than applying a uniform rule.

Table 5: **Effect of adaptive mask conditioning on COCO-Stuff with global captions**. t, ℓ, and h denote conditioning on timestep, layer, and attention head, respectively. Adaptive masking strongly outperforms hard masking; combining timestep and layer conditioning is crucial, while per-head conditioning provides further improvement.

Method	Conditioning			MIoU↑	Global quality				
	t	ℓ	h		FID↓	MAN-IQA↑	CLIP↑	Pick↑	IR↑
Hard masking				15.08	37.77	55.0	27.4	20.10	-0.87
Full	✓	✓	✓	23.37	29.89	56.0	29.7	21.62	0.14
Timestep only	✓			12.44	36.09	53.5	30.4	22.20	0.38
Layer only		✓		11.83	36.93	52.7	30.3	22.20	0.38
No per-head	✓	✓		17.22	32.63	53.3	29.7	21.58	0.07

4.3 Multi-conditioning

Fig. 1, Fig. 7 and Fig. 11 show qualitative results of SpaDiff combined with OminiControl, which the adapter design makes possible. The different modalities are incorporated successfully, with the generated images respecting all inputs in both in-distribution (Fig. 11) and out-of-distribution (Fig. 1, Fig. 7) settings. Note that in the out-of-distribution setting the model generates high-quality images at varying large sizes, despite being trained only at 512x512. All input conditions in Fig. 1 and Fig. 7 are generated from real-life architectural projects. The OminiControl fill adapter allows placing new designs in existing settings, the canny adapter offers control over fine-grained details such as facade design, and depth maintains correct perspective and massing. These results show that the adapter design enables spatial control in a multi-conditioning setting, even on real-world inputs outside the training distribution, making it directly applicable to design workflows.

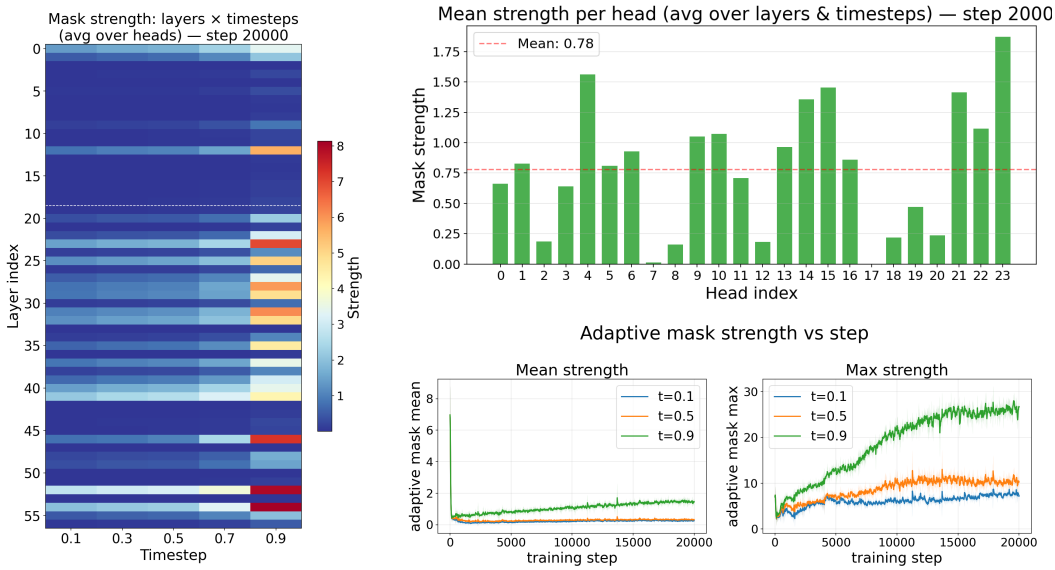

Figure 6: Insights into how layers, timesteps and attention heads contribute to the layout of the generated image, as learned in the adaptive mask. *Left*: the middle layers and some few final layers of the FLUX architecture influence the layout most. *Top right*: there is a significant difference in the contribution of different attention heads. *Bottom right*: Early denoising timesteps (t=0.9) contribute most to deciding the layout.

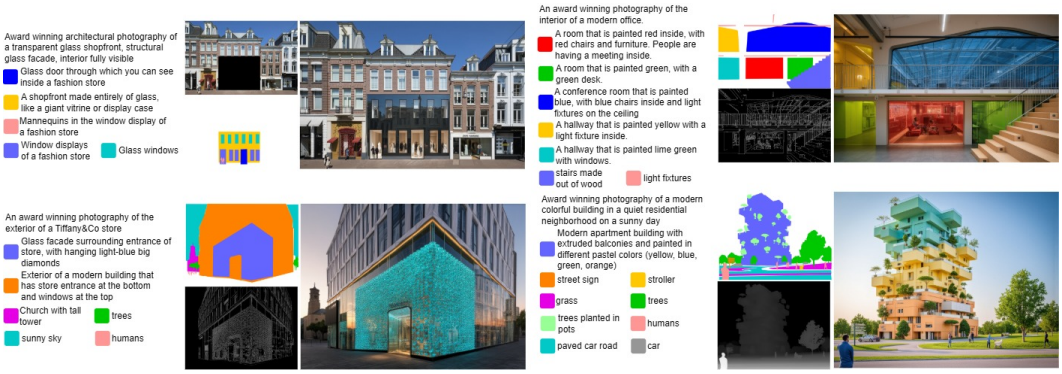

Figure 7: **Multi-conditioning on out-of-distribution architectural data.** SpaDiff is the segmentation adapter and Omini-Control is the fill, canny or depth adapter. Both modalities are successfully incorporated into the generated image.

5 Conclusion

SpaDiff achieves S2I performance on par with SOTA models while retaining an adapter design that combines with other spatial control modalities for fine-grained image generation control. The adaptive masking strategy alone enables open-vocabulary positional control in an adapter setting, with minimal parameter overhead, while conditional integration adds adherence to exact mask boundaries. Our study of decoupled conditional integration revealed ZeroScale as a useful mechanism to counteract the drawbacks of the rigid direct addition, at low additional computational cost.

SpaDiff's main limitations are resource constraints when generating large images with many densely described regions, and occasional attribute leakage between similarly shaped objects in out-of-distribution settings. When combined with adapters trained on less diverse data, the conditioning strength requires careful tuning.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing, designing or providing feedback on research methodology or experiments, implementing methods in code. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, cleaning and reformatting datasets, interpreting results. The remaining categories are not applicable to this work: generating synthetic data sets, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, assisting with translation, supporting qualitative and thematic data analysis. Additionally, we used generative AI tools for creating or editing software code, summarizing or analyzing existing literature and brainstorming. We have reviewed all AI-assisted work: LLM-generated code was verified and tested for correctness, AI-polished text was checked to ensure it has the intended meaning, suggestions regarding methodology and experiments were evaluated by the authors before being used and all summaries or explanations of existing literature were verified in the original source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. The SpaDiff architecture is fully described in Section 3, with the exact attention masking rules and the baseline configuration detailed in Section A. Training and evaluation settings are reported in Section 4. Anonymized source code is provided as supplementary material and covers the full pipeline: the model implementation, including the multi-conditioning setup combined with OminiControl; the data loaders for all datasets used in the paper, together with instructions for downloading and processing them; and the evaluation pipeline, with scripts for single-sample generation, full-dataset inference, and computation of every metric reported in the paper.

ACKNOWLEDGMENTS

Acknowledgments are withheld to preserve anonymity during review.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Black Forest Labs. FLUX.1. URL https://github.com/black-forest-labs/flux.

Danfeng Li, Hui Zhang, Sheng Wang, Jiacheng Li, and Zuxuan Wu. Seg2any: Open-set segmentation-mask-to-image generation with precise shape and semantic control. *arXiv preprint arXiv:2506.00596*, 2025.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet_plus_plus. In *European Conference on Computer Vision*, pp. 129–147. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, and Kwan-Yee K Wong. Place: Adaptive layout-semantic fusion for semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9264–9274, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14256–14266, 2023.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1191–1200, 2022.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Hong Zhang, Zhongjie Duan, Xingjun Wang, Yingda Chen, and Yu Zhang. Eligen: Entity-level controlled image generation with regional attention. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pp. 1–7, 2025a.

Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18487–18497, 2025b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in neural information processing systems*, 36:11127–11150, 2023.

Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16712–16722, 2025.

A ATTENTION MASKING IN OPEN-VOCABULARY S2I

The regional captions $p_1, \ldots, p_N$ are encoded with the same text encoder used for the global caption p_0, and their tokens are concatenated with the global text tokens and processed through the same text stream. To bind each regional caption to its corresponding image region, the text-image attention is modulated with a binary mask M whose log is added to the pre-softmax attention scores (Eq. (1)): an entry of 1 in M allows two tokens to interact, while an entry of 0 blocks them entirely. The mapping between regional text tokens and image tokens is determined by the segmentation map, and the mask follows three rules: (i) each regional text token interacts only with the image tokens in its corresponding region, (ii) image tokens within a region interact only with that region and with the background image tokens, and (iii) the global caption interacts only with background image tokens. Additionally, the conditional tokens are not allowed to interact with the text tokens, but are fully allowed to interact with each other and the image tokens.

Hard Masking Baseline Seg2Any (Li et al., 2025) applies these rules through Eq. (1) directly, where blocked entries contribute $\log(0) = -\infty$ to the attention scores. To balance regional separation with global coherence, the rules are applied strictly only in the middle transformer layers; in the remaining layers, the rules are relaxed so that image tokens can interact across regions and the global caption can attend to all image tokens. This approach requires fine-tuning the full FLUX backbone, since adding $-\infty$ to the scores of a frozen model disrupts the learned attention distributions and produces poor results. We adopt this scheme as a baseline by applying it directly to the frozen base model, without fine-tuning, in order to maintain our adapter design.

B ZEROSCALE ABLATIONS

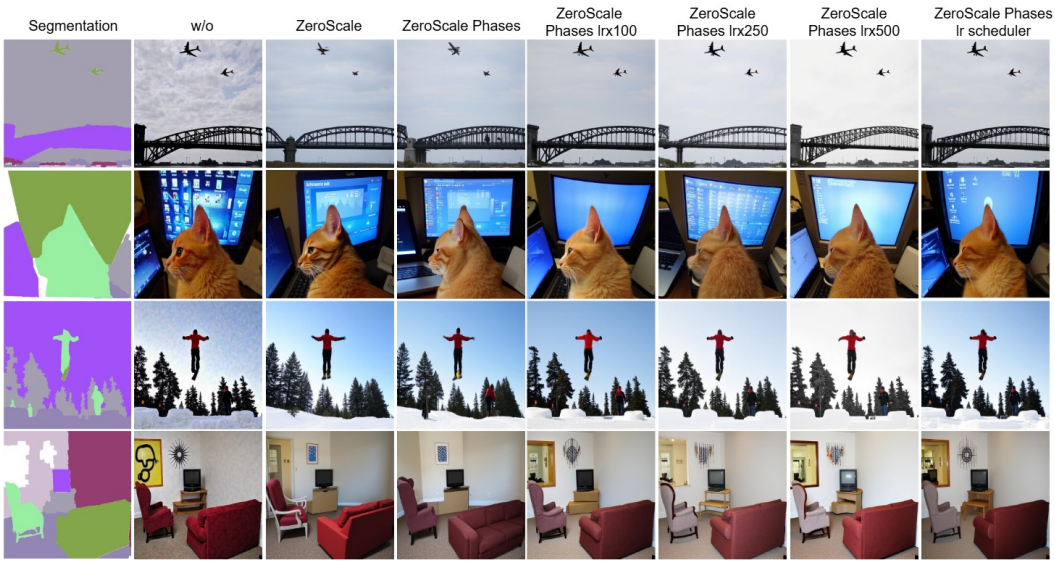

Figure 8: Visual results of ZeroScale ablations. Not using ZeroScale produces grainy images, which its naive usage manages to fix. Pre-training the decoupled stream and then jointly finetuning with a higher learning rate for the ZeroScale produces high-quality results more closely aligned with the input segmentation.

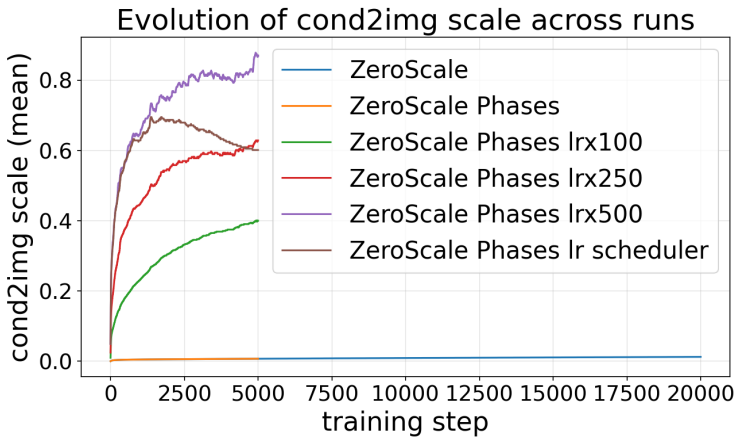

Figure 9: The evolution of ZeroScale parameters mean values durring training in different settings. While naive ZeroScale and training in phases with the same learning rate do not manage to open the gate enough, the more the learning rate is increased, the more the gate opens. Using a learning rate scheduler quickly opens the gate and then starts closing it, signaling that a fully open gate is not desirable.

C QUALITATIVE COMPARISON OF ADAPTIVE MASKING ABLATIONS

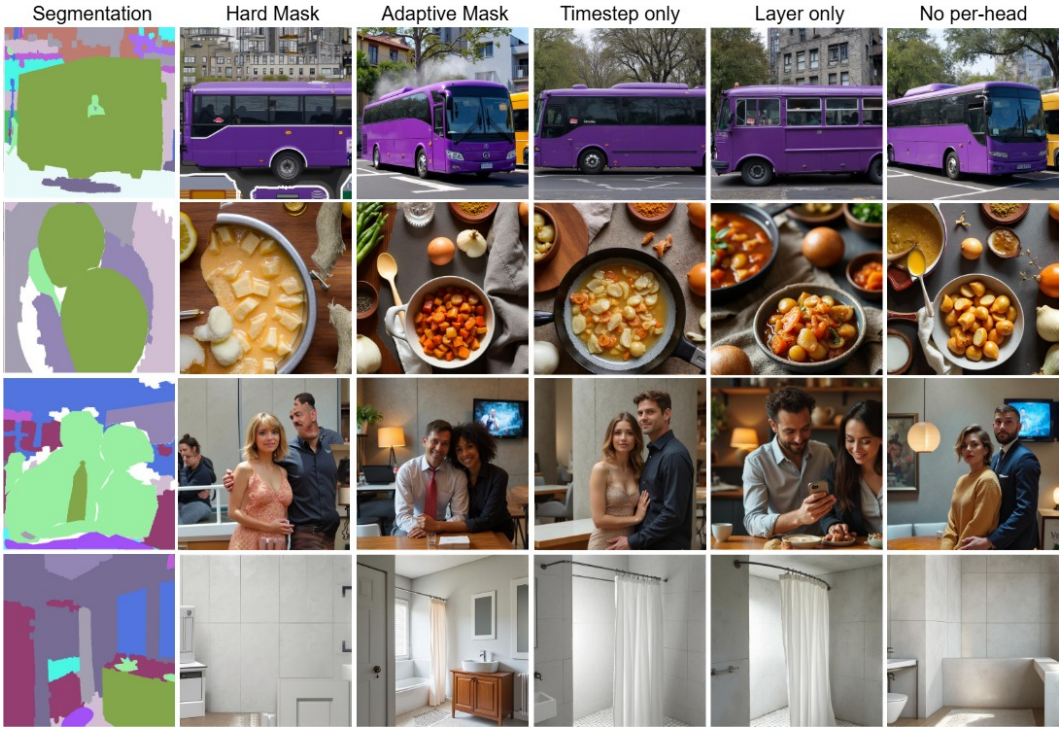

Figure 10: Visual results of Adaptive Masking ablations. Hard masks are worst performing, not managing to integrate the segmentation to produce a realistic image. Best performing is the full scheme of the Adaptive Mask and even though its ablations produce nice looking images, they only rarely and loosely respect the input layout.

D IN-DISTRIBUTION MULTI-CONDITIONING



Figure 11: Visual results of multi-conditioning with SpaDiff as the segmentation adaptor and Omini-Control as the canny adaptor. Both modalities are incorporated into the generated image and create results closer to the ground-truth then when used separately.