DISCONTROL: LOCAL CONTROL STRENGTH MODULATION VIA DISTORTION OF THE CONDITIONING SIGNAL

Anonymous authors
Paper under double-blind review

ABSTRACT

ControlNet has become the standard for spatial conditioning of diffusion models. At inference, a strength parameter modulates how strongly the condition is applied, trading adherence against diversity and image quality, and setting how much of the control map's detail is imposed, so that a coarse map can be used at a lower strength; the same knob can be set per region, so that different parts of an image follow the condition to different degrees. This is implemented by locally scaling the adapter's activations, an operation the network was never trained to receive, and we identify three limitations of it. First, under local modulation the scaled activations take the network out of distribution and image quality degrades severely. Second, a per-pixel strength map cannot be imposed on activations: it has to be resampled onto a representation coarser than the control map, as coarse as 8×8, and the backbone then mixes information between neighbouring positions, so a region boundary blurs either way and precise modulation on a specific object is out of reach. Third, the adherence actually realized at a given strength varies widely across images, as a result of the model picking up or not picking up on the control signal more randomly rather than smoothly. No method is explicitly designed for this modulation, and we propose a fundamental shift: from modulating internal activations at inference time to degrading the input at training time. In *DisControl*, the control signal is degraded by injecting noise at full pixel resolution and the adapter is trained on this family of degraded inputs, so that every strength is in distribution, region boundaries are sharp by construction, and modulation is part of the learned task rather than an intervention on top of it. Across backbones and control modalities, DisControl suppresses the dissimilarity artifacts that local modulation introduces, separates regions at pixel resolution, and reduces the variance of realized adherence. We show how our adapter can be trained from scratch or efficiently fine-tuned from existing ControlNets.

1 INTRODUCTION

Image diffusion models Ho et al. (2020); Song et al. (2020); Rombach et al. (2022); Saharia et al. (2022) generate high-quality visual content from text prompts, but many real-world use cases require a finer level of spatial control than the text modality alone can provide. ControlNet Zhang et al. (2023) is the dominant approach for conditioning a generative model on a pixel-aligned control map such as a depth map or an edge map. The condition is rarely applied at full strength: users tune an inference-time *conditioning scale*, or *strength*, which modulates adherence to the control map and trades it off against image quality, diversity, and how much detail the output inherits from the control signal. Intermediate values beyond full or no control are also important: a control map can guide scene layout or geometry coarsely without dictating its exact detail. The same knob can be set per region, taking high values where the condition must be followed closely and low values where the model should be left free, so that an object is reproduced faithfully while the rest of the image is only loosely guided. Both are implemented by linearly scaling the adapter's activations before they are added to the backbone, the local variant scaling them per region and relying on the spatial alignment between pixel space and the diffusion model's latent space.

This mechanism has two fundamental caveats. First, the activations it feeds to the backbone are out of distribution by construction, since during training the adapter is only ever added at full strength. Training on the same distribution of local strengths does not repair this: an adapter whose own

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.

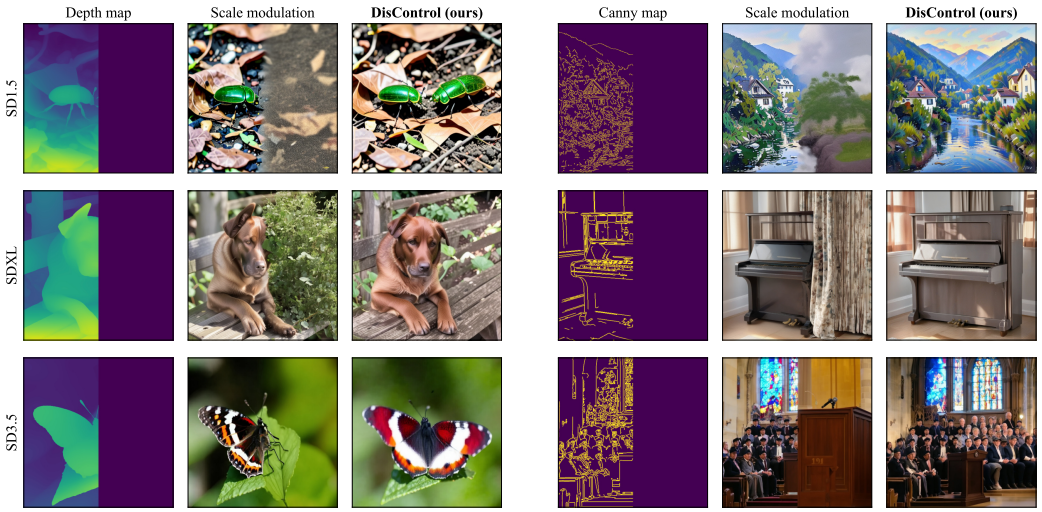

Figure 1: Under local modulation the activation-scaling baseline leaves a visible seam between the controlled and released halves, often changing scene, style or colour across the boundary, while DisControl keeps the image coherent.

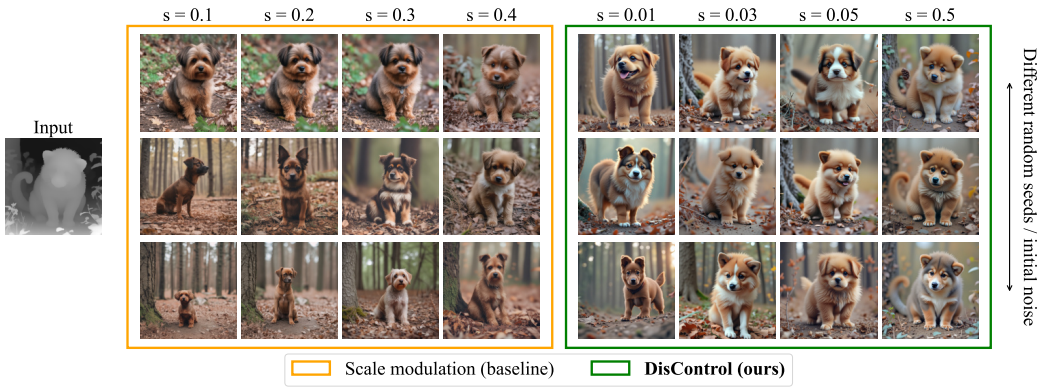

Figure 2: Activation scaling gives an unreliable strength knob. At a fixed strength value the realized adherence to the control map varies wildly across images, while DisControl is more consistent.

activations are scaled minimises the diffusion objective most easily by becoming insensitive to their magnitude, so its entire response collapses into the bottom fifth of the knob and modulation is lost rather than learned (Section 5.3). Second, a per-pixel strength map cannot be imposed on internal activations. It has to be resampled onto the representation the residuals are added to, which is always coarser than the control map and can be as coarse as 8×8, and the backbone might also mix information between neighbouring positions, so a boundary blurs further however fine that representation is.

We hypothesise that these two caveats are the reason for the two failure modes we observe under local modulation. First, when two regions are given very different strengths, the model produces two disjoint regions that do not correspond to each other (Figure 1, Table 1). Second, at the region boundary it cannot hold high adherence on one side and low adherence on the other at fine resolution (Table 2). We observe a third failure mode, this one independent of locality: at a given strength value the realized adherence varies widely across images, as if the model either picks up the control signal or does not (Figure 2). A global scale below 1 leaves the activations as unseen in training as a local one does, so the first caveat is the natural explanation here too, although we do not test it.

Work on control strength has largely targeted how much of it to apply, selecting the scale automatically or predicting where the condition should be trusted, and leaves the mechanism itself untouched: the strength is still realized by scaling activations at inference, so the two caveats above are inherited in full.

We propose a novel mechanism of strength modulation, *DisControl*, which relies on degrading the control signal at the pixel level during training and inference time, as opposed to relying on an inference-time-only activation interpolation mechanism which is by construction out of distribution and operates at lower resolution. Strength is lowered by blending noise into the control map itself at the pixel level, so the modulation can take place at the pixel resolution, instead of the latent space lower resolution which could additionally have cross-token information leakage. The adapter is trained on a wide and representative distribution of different strength values at different regions, and at inference provides reliable behaviour across them, which leads to higher quality and cross-region consistency.

Against a separately trained vanilla ControlNet driven by the standard scale knob, on three backbones (SD1.5, SDXL, SD3.5), three control modalities (depth, canny edges and surface normals) and two test sets, DisControl suppresses the dissimilarity artifacts that local modulation introduces, separates regions at pixel resolution, and lowers the variance of realized adherence.

In summary, our contributions are:

- We identify two caveats of activation-scaling strength modulation, out-of-distribution activations and the impossibility of imposing a pixel-resolution strength map on internal activations, and quantify the three specific failure modes they produce.
- We propose DisControl, which modulates control strength by degrading the input control signal at pixel resolution and trains the adapter on that family of inputs, so that every strength is in distribution and local boundaries are sharp by construction.
- We validate all three results against a separately trained vanilla ControlNet across three backbones, three control modalities and two test sets, and show that a public ControlNet can be adapted into the DisControl layout and reach its adherence plateau in a small fraction of the steps a from-scratch run needs.

2 Related work

Text-to-image diffusion models Rombach et al. (2022); Podell et al. (2023); Esser et al. (2024); Saharia et al. (2022) offer limited means of spatial control through language alone, owing to the inherent ambiguity of text descriptions and the constraints of text encoders Chen et al. (2020); Oquab et al. (2023); El Banani et al. (2024). Classifier-free guidance Ho & Salimans (2022) enables an inference-time balance between sample quality, diversity, and control fidelity, yet does not address fine-grained spatial conditioning. Cross-attention manipulation Hertz et al. (2022) and instruction-following editing Brooks et al. (2023) provide indirect, prompt-driven means of control but do not consume a dense spatial signal. Earlier approaches enrich generation with auxiliary signals such as edge maps, sketches, or low-resolution images Li et al. (2022); Mou et al. (2024); Voynov et al. (2023), while another line grounds generation on bounding boxes or layouts Li et al. (2023); Balaji et al. (2022). ControlNet Zhang et al. (2023) introduces a lightweight adapter that conditions a frozen pretrained diffusion model on a pixel-aligned control signal, substantially improving spatial controllability. This seminal work became the de-facto standard for spatial conditioning in diffusion models, and has been widely adopted in practical applications. It also introduced control modulation by means of activation scaling, as described in Section 3. Subsequent work reduces parameter or compute overhead Fang et al. (2025); Peng et al. (2024); Wang et al. (2024); Mo et al. (2024), composes multiple conditions Huang et al. (2023), extends to image-prompt conditioning Ye et al. (2023). Recently, Li et al. (2024) showed significant improvements in control adherence and efficiency, and we use this as additional baseline for comparison. However, all the aforementioned improvements on ControlNet still rely on the same activation-scaling mechanism for control modulation and suffer from the same limitations. SDEdit Meng et al. (2021) is a pixel-space diffusion model, and showed the usefulness of beginning the diffusion process from a partially noisy image, as opposed to pure noise, effectively modulating the similarity between the generated image and the starting reference image. However, the approach is restricted to RGB modality-only and in pixel-space diffusion models. Instead, DisControl consists

in conditioning the generative process to a partially noisy map, which is applicable in any generative paradigm, modality and diffusion space. Recently, other parameter-efficient adapters such as LoRA Hu et al. (2022) have been used to fine-tune diffusion model for customization Ruiz et al. (2022) as well as spatial control Xu et al. (2024).

LocalControl Zhao et al. (2023) is concerned with ensuring semantic coherence between a locally applied control map and the global text prompt, not with the reliability or resolution of strength modulation. SmartControl Liu et al. (2024) automatically adjusts the strength map to suppress quality degradation caused by inconsistencies between the depth map and the text prompt. Its strength map is an internal correction signal, not a user-controllable input, and its goal is prompt-condition mismatch recovery rather than reliable user-driven spatial control. DC-ControlNet Yang et al. (2025) deals with the composition of multiple control modalities across disjoint image regions based on bounding boxes, but does not support per-region modulation of control strength. The majority of previous work relies on the paradigm of activation-interpolation mechanism that we describe in Section 3.

We use as baseline of strength modulation the standard ControlNet scale parameter. It is the default strength modulation interface exposed by widely-used libraries such as Diffusers von Platen et al. (2022) and ComfyUI, and remains the standard in practice across all modern backbone architectures. Moreover, we compare more recent improvements of the ControlNet architectures, including ControlNet++ Li et al. (2024) and SmartControl Liu et al. (2024), but they both still rely on the same activation-scaling mechanism for strength modulation, and thus suffer from the same limitations as the original ControlNet.

3 Background

We refer to the text-only model as the unconditional base model, and define its hidden features at layer l as $\mathbf{h}_l^u \in \mathbb{R}^{C_l \times H_l \times W_l}$. ControlNet model freezes the weights of the base model and trains a smaller adapter module, which takes as input the control signal such as depth or edge map $\mathbf{c} \in \mathbb{R}^{H \times W}$, and generates conditional activations at different layers, defined as $\mathbf{h}_l^c \in \mathbb{R}^{C_l \times H_l \times W_l}$. Finally, the conditional activations are added to the unconditional activations, obtaining $\mathbf{h}_l = \mathbf{h}_l^u + \mathbf{h}_l^c$, which are used to generate the final model prediction. The adapter module is trained with the usual diffusion objective and can then sample from the conditional distribution $p(x|c)$.

During inference, the user can modulate the level of adherence to the control signal by scaling the conditional activations with a scalar $s \in [0, 1]$, obtaining $\mathbf{h}_l = \mathbf{h}_l^u + s \cdot \mathbf{h}_l^c$. For $s = 0.0$, the model samples from the unconditional distribution and for $s = 1.0$ from the conditional distribution. Empirically, intermediate values of s can be used to smoothly modulate the control strength between the two distributions. This approach loosely resembles the classifier-free guidance (CFG) approach, where the model samples from the distribution $p_s(x) \propto p(x|c)^s \cdot p(x)^{1-s}$.

This approach has also been used for locally modulating the strength of the control signal, by replacing the scalar s by a strength map $s \in [0, 1]^{H \times W}$, where s_{ij} is the desired level of adherence to the control map at pixel (i, j). One of the limitations is that it requires to downsample the strength map at each latent feature resolution, reaching resolutions as low as 8×8 at the bottleneck in UNet models. Moreover, it relies on the assumption that the latent space and feature space are spatially aligned with the pixel space, which appears to be particularly problematic for the transformer-based model SD3.5.

4 Method

As opposed to previous methods which modulate control by scaling internal activations, DisControl achieves strength modulation by degrading the input control signal before it enters the network. The main property of our method is that the information contained in the control signal can be continuously degraded by varying the amount of noise injected into the control signal, smoothly transitioning between fully conditional (no noise) and unconditional (all noise) generation. Moreover, because modulation is applied to the input rather than internal activations, DisControl paradigm is model-agnostic and can be applied to different models and adapters with no modifications.

Input degradation. DisControl can use different mechanisms to degrade the signal, but we empirically find uniform noise to work well. That is, we construct the degraded control signal $\tilde{c}$ as a convex

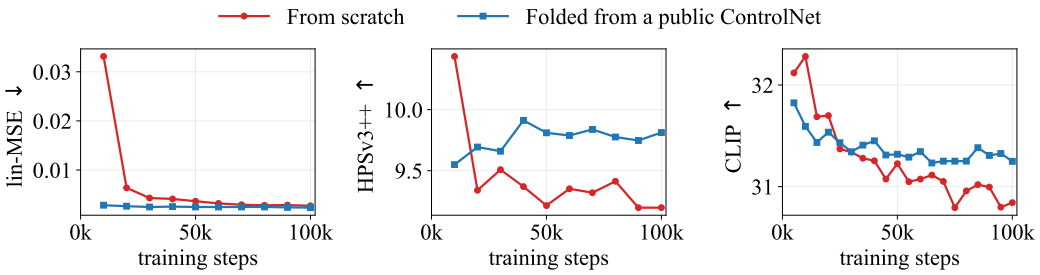

Figure 3: Folding a public ControlNet into the DisControl layout reaches the adherence plateau at the first checkpoint we evaluate, where training from scratch needs tens of thousands of steps, and it ends ahead on image quality and text alignment as well. SD1.5 depth.

combination of the original control signal c and a noise map $\eta \sim \mathcal{U}(0,1)^{H\times W}$: $\tilde{c} = S\odot c+(1-S)\odot\eta$, where $\odot$ denotes element-wise multiplication.

Setting $S = s \cdot \mathbf{1}$ for a scalar $s \in [0,1]$ recovers global control with uniform strength across the image. The key advantage of our formulation is that S can take arbitrary spatially varying values, enabling per-pixel strength modulation at the full resolution of the control map, without any resolution bottleneck. This stands in contrast to activation-based local control, which is constrained by the resolution of the UNet's internal feature maps and relies on a weak spatial alignment assumption between activation and pixel space.

Training. The adapter is trained on the same family of inputs it is asked to handle at inference. For each training image we draw a strength map S, and the network receives $\tilde{c}$ and S concatenated along the channel dimension, so that it can tell a map degraded by noise blending from a clean map of a genuinely grainy texture, which is legitimately noisy. Half of the maps are built from a segmentation mask of an object in the image, randomly perturbed, with one strength inside it and another outside; of the remainder, one in five is constant over the whole image and the rest partition it into two regions of a random shape (vertical and horizontal halves, rectangles, ellipses, circles and their centered variants), again with one strength per region. Strength values are drawn from a mixture placing 10% of its mass at $s = 0$, 20% at $s = 1$, 45% on $\mathcal{U}(0, 0.2)$ and 25% on $\mathcal{U}(0.2, 1)$. Any strength map a user supplies at inference, a global scalar, a bisection or an object mask, is therefore a sample from the distribution the adapter was trained on, and no input it receives at inference is one it has never seen. This is the property activation scaling cannot have, since its modulation exists only at inference time.

Adapting an existing ControlNet. DisControl does not have to be trained from scratch. A public ControlNet can be folded into the two-channel layout directly: the control filter of the first convolution is the sum of its pretrained counterparts and the strength filter is initialised to zero, so at initialisation the adapter reproduces the public model exactly and training only has to learn the response to the strength map. Figure 3 compares a model initialised this way against the same recipe trained from scratch. The adapted model is at its adherence plateau by the first checkpoint we evaluate, where the from-scratch run needs tens of thousands of steps to arrive. At its earliest checkpoints the from-scratch run scores higher on image quality and text alignment, while its adherence is still far from converged; from 25k steps on, the adapted model leads on both and the gap grows rather than closes as training proceeds.

5 Experiments

5.1 Experimental Setup

Models, data and training. We evaluate on three backbones spanning two architectures, Stable Diffusion 1.5 and SDXL 1.0 (UNet Ronneberger et al. (2015); Podell et al. (2023)) and Stable Diffusion 3.5-Medium (transformer Peebles & Xie (2023); Esser et al. (2024)), and on three control modalities: depth maps extracted with Depth-Anything-V2-Large Yang et al. (2024), Canny edge

maps, and surface normals extracted with DSINE Bae & Davison (2024). All adapters are ControlNets trained on a 982,734-image subset of the MONET corpus Aubin et al. (2026) at the backbone's native resolution, 512×512 for SD1.5 and 1024×1024 for SDXL and SD3.5. We report SD1.5 and SDXL at 100k training steps and SD3.5 at 60k, fixed in advance rather than selected on the test metrics. Full hyperparameters are in the supplementary material.

Protocol. For every (backbone, modality) pair we train *two* adapters on identical data, recipe and number of steps: a vanilla ControlNet, evaluated with the standard activation-scaling knob, and DisControl, evaluated with noise modulation. All comparisons are paired: both arms are evaluated on the same 1000 images from the held-out MONET test split, disjoint from training, with a per-image seed shared across arms and byte-identical region masks. We use three region settings. **Object masks** obtained with SAM Kirillov et al. (2023), with strength 1 inside the largest object and 0 outside, are the main local setting. A fixed vertical **bisection** is the most extreme local case and the one in which cross-region coherence can be measured directly. A single **global** strength per image is used for the variance analysis.

Metrics. Image quality is measured with HPSv3++ Liu et al. (2026) over the whole image. Coherence across the region boundary is measured with DreamSim Fu et al. (2023) between the two halves of the image. Adherence to the control signal is measured between the input control map and the map re-extracted from the generated image: for depth, the residual MSE after fitting an optimal scale and shift, $\min_{a,b}\mathrm{MSE}(d_1, a\cdot d_2+b)$, which is invariant to the global scale ambiguity of monocular depth and has markedly lower per-image variance than plain MSE; for Canny, distances between binary edge maps; for surface normals, the mean angular error in degrees.

Second test set. To test out-of-domain generalisation we repeat the object-mask setting on 1000 images of COCO val2017 Lin et al. (2014), using the largest non-crowd instance annotation as the region mask.

5.2 Quality under Local Modulation

Both arms receive the same binary strength map, one half of the image at $s=1$ and the other at $s=0$. This is the most extreme local setting and the one in which the two regions can be compared against each other directly. Under activation scaling the released half frequently departs from the scene, style and palette of the controlled one, leaving a visible seam. Table 1 reports HPSv3++ over the whole image and DreamSim between the two halves, on 1000 paired images per cell. On both UNet backbones DisControl improves HPSv3++ in all six cells, by 14.8 to 99.4% relative and on 68 to 87% of individual images; DreamSim falls in all nine cells, by 9.0 to 27.7%.

The result also holds on natural images. On 1000 COCO val2017 photographs, with COCO instance annotations as the region masks, all six SD1.5 and SDXL cells favour DisControl, by 16.3 to 59.3% HPSv3++ on 62 to 80% of images, and CLIP similarity, which favours the scale baseline in every cell of the synthetic set, reverses in all six (Table 7). Table 6 reports the same comparison on SAM object masks rather than a bisection.

SD3.5 does not follow this pattern on HPSv3++. Its deltas are -1.6% (depth), -2.3% (canny) and -14.1% (normals) on the bisection, and -11.6, -9.5 and -14.9% on object masks. On depth and canny DisControl is preferred on 47% of images in both cells, a near-even split; normals is the only reversal. Two factors bear on this comparison. First, we hypothesise that the two arms are not on equal footing at $s=0$: a ControlNet adapter generally imposes some quality cost on its backbone, and activation scaling removes the adapter from the released region entirely, so that region is generated by the backbone alone, whereas noise modulation keeps the adapter active across the whole image. Second, we hypothesise that HPSv3++ does not fully capture the failure mode at issue, a dissimilarity between the two regions rather than a degradation of the image as a whole. Figure 4 shows two cases in which it ranks the baseline above DisControl while the baseline output is split across the strength boundary. DreamSim, which measures that dissimilarity directly, favours DisControl on all three SD3.5 modalities, falling by 26.5, 22.0 and 9.5% on 85, 81 and 67% of images. The boundary measurement of Section 5.3 is consistent, with SD3.5 showing the widest margin of any backbone.

		HPSv3++ ↑				DreamSim ↓			
Backbone	Modality	base	ours	Δ%	win%	base	ours	Δ%	win%
SD1.5	depth	4.641	7.338	**+58.1**	85	0.479	0.352	**−26.6**	86
	canny	4.401	6.934	**+57.5**	86	0.472	0.351	**−25.5**	84
	normals	5.516	7.335	**+33.0**	76	0.439	0.347	**−20.9**	81
SDXL	depth	7.847	9.010	**+14.8**	68	0.400	0.308	**−23.0**	84
	canny	5.169	8.550	**+65.4**	87	0.460	0.333	**−27.7**	87
	normals	3.138	6.256	**+99.4**	83	0.435	0.396	**−9.0**	63
SD3.5	depth	10.187	10.020	−1.6	47	0.423	0.311	**−26.5**	85
	canny	8.765	8.567	−2.3	47	0.445	0.347	**−22.0**	81
	normals	10.187	8.751	−14.1	28	0.398	0.361	**−9.5**	67
Average		*6.650*	*8.085*	***+21.6***	*67*	*0.439*	*0.345*	***−21.4***	*80*

Table 1: Under split-half modulation DisControl improves image quality in every cell of both UNet backbones and reduces the dissimilarity between the two halves in all nine cells. SD3.5 is the exception on quality but not on dissimilarity. Win% is the share of the 1000 paired images on which DisControl is better.

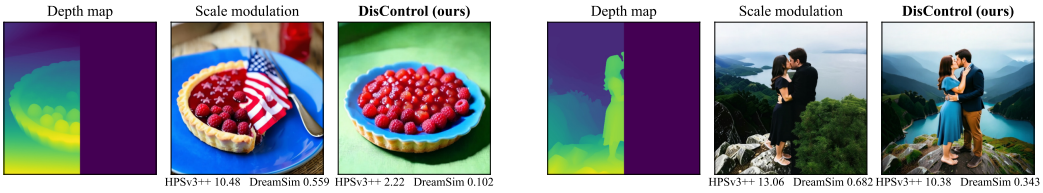

Figure 4: On SD3.5 the baseline splits the image into two unrelated halves at the strength boundary while DisControl stays coherent, yet HPSv3++ prefers the baseline in both. Chosen among the pairs where the two metrics disagree. Depth, left half $s{=}1$, right half $s{=}0$.

5.3 Resolution at the Region Boundary

This experiment measures the spatial resolution at which a boundary in the strength map is actually realized in the output. Both arms receive the same object mask, strength 1 inside and 0 outside. We measure adherence separately on the two sides of the boundary, restricted to the band of grid patches the boundary passes through: dividing the image into an $N \times N$ grid, a larger N gives a narrower band and therefore a finer localisation of where the modulation takes effect. We report the separation ratio, adherence outside the mask over adherence inside it; a value of 1 means the two regions were treated identically and larger is better.

Table 2 reports three backbones, three modalities and four band widths. DisControl separates the regions better in 33 of the 36 cells, and the margin is larger at the narrowest band than at the coarsest, $+13.5\%$ on average at $N{=}16$ against $+18.5\%$ at $N{=}128$, consistent with the baseline's mask being resampled onto feature maps that cannot represent it. The three exceptions are canny at the coarsest bands, where the two arms are within a few percent of each other.

The largest margins are on SD3.5, and they increase with band resolution rather than decreasing: $+17.5\%$ at $N{=}16$ rising monotonically to $+47.7\%$ at $N{=}128$, the largest single gain in the table, on the backbone whose quality result is least favourable to us. This might be due to the transformer architecture leaking information about the control map across tokens. It also shows that this limitation is not confined to smaller or less capable models; here the newest backbone is the most affected.

A scale baseline trained on DisControl's own distribution of per-pixel strength maps reaches a comparable separation (Section 5.5), but it does not support smooth interpolation: its adherence is already saturated at $s{=}0.2$ and flat from there to 1, so the knob it exposes is effectively a switch (Figure 5).

		separation ratio, outside / inside ↑											
		$N = 16$			$N = 32$			$N = 64$			$N = 128$		
Backbone	Modality	base	ours	Δ%	base	ours	Δ%	base	ours	Δ%	base	ours	Δ%
SD1.5	depth	1.7	**2.6**	**+50.2**	1.4	**2.2**	**+54.8**	1.2	**1.8**	**+53.2**	1.0	**1.4**	**+41.2**
	canny	**2.0**	2.0	−1.4	1.9	**2.0**	**+5.2**	1.8	**2.0**	**+12.4**	1.6	**2.0**	**+21.9**
	normals	1.4	**1.7**	**+23.7**	1.3	**1.6**	**+19.0**	1.2	**1.4**	**+12.7**	1.1	**1.2**	**+10.4**
SDXL	depth	2.9	**3.2**	**+11.9**	2.5	**3.0**	**+20.1**	2.0	**2.5**	**+23.8**	1.6	**1.9**	**+20.1**
	canny	**2.3**	2.1	−6.5	**2.2**	2.2	−1.4	2.1	**2.2**	**+4.4**	2.0	**2.2**	**+8.7**
	normals	1.8	**2.1**	**+19.7**	1.7	**2.0**	**+17.5**	1.6	**1.8**	**+15.4**	1.4	**1.5**	**+10.0**
SD3.5	depth	2.4	**2.8**	**+17.5**	2.0	**2.5**	**+25.1**	1.7	**2.2**	**+33.1**	1.3	**1.9**	**+47.7**
	canny	2.1	**2.2**	**+2.6**	2.1	**2.2**	**+3.6**	2.1	**2.1**	**+4.4**	1.9	**2.0**	**+7.2**
	normals	1.2	**1.5**	**+25.4**	1.2	**1.5**	**+21.7**	1.2	**1.4**	**+18.6**	1.1	**1.3**	**+15.1**
Average		2.0	**2.2**	**+13.5**	1.8	**2.1**	**+17.8**	1.7	**1.9**	**+16.8**	1.4	**1.7**	**+18.5**

Table 2: DisControl separates the controlled from the released region more sharply in 33 of the 36 cells, and its margin widens as the boundary band narrows. The three exceptions are canny at the coarsest bands. Larger is better.

	per-image spread σ at mean adherence μ, lin-MSE								
	$\mu = 0.006$			$\mu = 0.020$			$\mu = 0.040$		
Backbone (depth)	base	ours	Δ%	base	ours	Δ%	base	ours	Δ%
SD1.5	0.0053	0.0047	**−12.8**	0.0109	0.0097	**−11.6**	0.0165	0.0147	**−10.9**
SDXL	0.0058	0.0056	**−4.8**	0.0105	0.0099	**−6.4**	0.0148	0.0137	**−7.3**
SD3.5[†]	0.0062	0.0049	**−20.4**	0.0109	0.0097	**−11.5**	0.0152	0.0143	**−6.0**

Table 3: Per-image spread of adherence at matched mean adherence, on depth. Each entry is the fitted $\sigma = k\,\mu^{\gamma}$ read at that μ; negative Δ favours DisControl, whose spread is lower at every level on all three backbones. [†]Evaluated at 100k training steps rather than the 60k used for every other SD3.5 result; the run had converged at 60k.

5.4 Variance of Realized Adherence

Figure 2 shows that under activation scaling the adherence realized at a single strength value ranges from near-perfect to near-absent across images. The two methods cannot be compared at equal strength, since a given value is a different operating point for each of them. We compare them at matched *realized adherence* instead: we sample one strength per image over the whole range, fit the mean and the spread of adherence jointly on the raw per-image measurements by heteroscedastic maximum likelihood, and read off the spread at a given mean. A method is better if, at the adherence level requested, the images it produces are more tightly grouped around it.

Table 3 reports depth for all three backbones at three adherence levels spanning the useful range. DisControl's spread is lower at every level on every backbone, by 12.8 to 10.9% on SD1.5, 4.8 to 7.3% on SDXL and 20.4 to 6.0% on SD3.5, where the margin is largest under tight control and narrows as the constraint loosens.

We observe the same behaviour across different backbones and modalities; the per-modality detail is reported in the supplementary material.

5.5 Comparison to Baselines

Table 4 compares DisControl against five alternatives on SD1.5 depth, 1000 images per cell, reporting image quality and the OFF/ON ratio of adherence between the released and the controlled region, for which a larger value means the mask was honoured more strictly.

Beyond the standard scale baseline used throughout the paper, each of the remaining four fails for a reason that follows from its mechanism.

Method	SAM masks		split-half	
	HPSv3++	OFF/ON	HPSv3++	OFF/ON
DisControl (ours)	**6.983**	4.059	**7.338**	**4.198**
Vanilla ControlNet, scale	4.963	3.146	4.641	3.389
Trained with intermediate scale	4.981	**4.850**	4.913	3.161
Stock ControlNet + noised input	1.096	2.539	0.554	1.557
ControlNet++	5.298	2.303	5.845	2.556
SmartControl (RV5.1 base)	7.503	1.759	8.242	2.002

Table 4: Training activation scaling with intermediate scales does not close the quality gap, 4.981 against our 6.983 on SAM masks. SmartControl is excluded from the bolding because its quality is measured on a different base model. Larger is better throughout. SD1.5 depth.

ControlNet++ Li et al. (2024) improves the adapter and its training, but modulation is still applied by scaling activations, so local modulation takes it out of distribution exactly as it does the standard baseline. **SmartControl** Liu et al. (2024) does not accept a strength map: it predicts one internally from the control map and the text prompt. We therefore give it the closest equivalent input, a control map masked over the released region, and it neither holds the strength where it is asked for nor leaves the masked region free, filling it in instead. **Stock ControlNet fed our degraded control map**, without retraining, follows the noise in the map as if it were signal, which degrades the output. **Activation scaling trained with intermediate scales** receives the training signal DisControl receives, but scaling remains its mechanism: it becomes insensitive to activation magnitude and saturates, as described in Section 5.3, so it separates a binary mask without interpolating between strengths.

6 Conclusion

The conditioning scale is the knob every ControlNet deployment relies on, and it does not do what its users assume. The adherence it delivers at a fixed value varies widely from image to image, so it is not a usable proxy for how closely the output will follow the control map; and set per region, it cannot hold a sharp boundary between a controlled and a released area, because the mask has to be resampled onto feature maps far coarser than the image. Underneath both is the same fact: scaling activations is an intervention applied at inference to activations the network was never trained to receive, and it acts at the resolution of those activations rather than at the resolution of the control signal.

We proposed DisControl, which moves the modulation to the input. The control signal itself is degraded under a per-pixel strength map at full resolution, and the adapter is trained across the resulting family of inputs, so every strength is in distribution and region boundaries are as sharp as the map that defines them. Across three backbones and three modalities, this lowers the variance of realized adherence at matched adherence, separates regions better at every band width we measure, and improves image quality under local modulation on both UNet backbones. On SD3.5 the whole-image preference score reverses on normals and is close to a draw on depth and canny, while the dissimilarity between the modulated and released regions still falls on all three of its modalities, as does its boundary separation.

The failure we describe is not specific to ControlNet. Classifier-free guidance Ho & Salimans (2022) interpolates between two sets of activations at inference in exactly the same way, and whether it suffers the same unreliability, and whether input-space degradation offers a way around it, is an open question this analysis raises.

Use of Large Language Models. Large language models were used throughout the preparation of this work. They assisted in implementing the method and the training, inference and evaluation code, and in the data processing that produces the control maps; in analysing and interpreting the experimental results; in producing the figures and the scripts that generate them; in locating and formatting references and summarising related work; and in revising the text of this paper. The research questions, the experimental design and the protocol were determined by the authors. The authors reviewed all AI-assisted output and take full responsibility for the contents of this submission, including every reported number and claim.

REFERENCES

Benjamin Aubin, Gonzalo Iñaki Quintana, Onur Tasar, Sanjeev Sreetharan, Urszula Czerwinska, Damien Henry, and Clément Chadebec. Monet: A massive, open, non-redundant and enriched text-to-image dataset. *arXiv preprint arXiv:2605.21272*, 2026.

Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

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

Zheng Fang, Lichuan Xiang, Xu Cai, Kaicheng Zhou, and Hongkai Wen. Flexcontrol: Computation-aware controlnet with differentiable router for text-to-image generation. *arXiv preprint arXiv:2502.10451*, 2025.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv. org/abs/2208.01626*, 3:3, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

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.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

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.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22511–22521, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Yijun Liu, Jie Huang, Zeyue Xue, Yuming Li, Ruizhe He, Haoran Li, Shijia Ge, and Siming Fu. Hpsv3++: Scaling reward models across the full spectrum of diffusion model capabilities. *arXiv preprint arXiv:2606.14657*, 2026.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7465–7475, 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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Cong Wang, Jiaxi Gu, Panwen Hu, Haoyu Zhao, Yuanfan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. Easycontrol: Transfer controlnet to video diffusion for controllable generation and interpolation. *arXiv preprint arXiv:2408.13005*, 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Yifeng Xu, Zhenliang He, Shiguang Shan, and Xilin Chen. Ctrlora: An extensible and efficient framework for controllable image generation. *arXiv preprint arXiv:2410.09400*, 2024.

Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. Dc-controlnet: Decoupling inter-and intra-element conditions in image generation with diffusion models. *arXiv preprint arXiv:2502.14779*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

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.

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.

Yibo Zhao, Liang Peng, Yang Yang, Zekai Luo, Hengjia Li, Yao Chen, Zheng Yang, Xiaofei He, Wei Zhao, Boxi Wu, et al. Local conditional controlling for text-to-image diffusion models. *arXiv preprint arXiv:2312.08768*, 2023.

A Perceptual Interpretation of the Strength Parameter

Natural images have most of their energy concentrated at low frequencies. Degrading a control map with spectrally flat noise therefore removes high-frequency detail first, gradually erasing coarser structure as noise increases. This gives the strength parameter an intuitive perceptual meaning: reducing strength first softens fine details of the control map before progressively removing its global structure, providing a predictable and interpretable modulation trajectory.

B Numerical Variance Comparison

Figure 6 shows the mapping from the strength knob to mean adherence, and the per-image spread against that mean, for depth on all three backbones. Table 5 reports the same measurement on the remaining modalities.



Figure 5: Training an activation-scaling adapter on the per-pixel strength distribution compresses its whole response into the bottom of the knob: the lower row is already adherent at s=0.1 and flat after it, while the baseline above is still far from the control map there. One example, SD1.5 depth; lin-MSE under each panel.

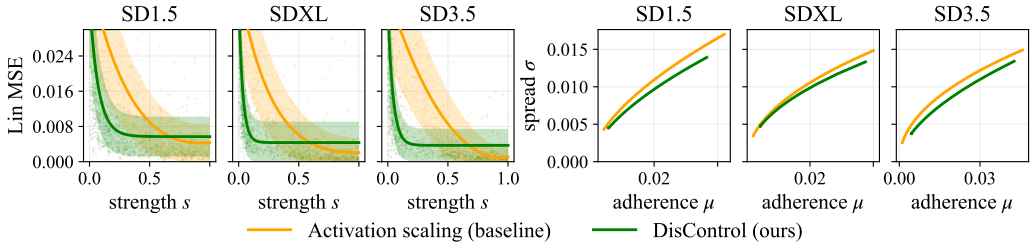

Figure 6: The variance result in fitted form: the baseline's spread is the wider one at every level, on all three backbones. Left, every image against the strength it was given, with the fitted mean and a $\pm 1\sigma$ band; right, spread against mean adherence, measured (solid) with the fit of Table 3 dashed through it. Depth, 1000 images per arm.

C Extrapolation to hand-drawn regions

Our model has been trained on a diverse set of control regions for local control, including simple geometric shapes and SAM masks, augmented with random geometric transforms and morphological operations. Figures 22, 23, and 24 show qualitative examples of the model's ability to generalize to hand-drawn regions.

D Qualitative Examples

We provide several examples in Figures 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18 for local and global control modulation across SD1.5, SDXL, and SD3.5 and depth and canny modalities. Figures 19, 20, and 21 show examples on COCO.

E Training Hyperparameters

All runs are trained for 100k steps with a learning rate of 10^{-5} and an effective batch size of 4; we report SD1.5 and SDXL at 100k steps and SD3.5 at 60k. SD3.5 uses a constant schedule with a 1000-step linear warmup, SD1.5 and SDXL a constant schedule. The backbone is frozen; only the ControlNet weights are updated. We use AdamW with default $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 10^{-2}, in its 8-bit form for SDXL.

The strength map encoder is a single Conv2d layer (no activation), trained from scratch jointly with the rest of the ControlNet. For SD3.5, where the control signal is VAE-encoded before being passed

		per-image spread σ at mean adherence μ								
		tight			mid			loose		
Modality	Metric	base	ours	$\Delta\%$	base	ours	$\Delta\%$	base	ours	$\Delta\%$
canny	Chamfer (px)	0.93	1.00	$+8.0$	1.60	1.55	-2.8	2.19	2.00	$\mathbf{-8.6}$
normals	degrees	3.47	4.70	$+35.6$	5.02	6.05	$+20.5$	7.28	7.80	$+7.1$

Table 5: Beyond depth the picture is mixed. On canny DisControl is tighter under loose control and wider under tight control; on normals it is wider across the range. Negative Δ favours DisControl. SD1.5; levels are $\mu = 2/4/6$ for canny and $5/10/20$ for normals.

to the ControlNet, the Conv2d uses a kernel size and stride both equal to the VAE patch size, so the encoder output matches the spatial dimensions of the latent and can be added elementwise. For SD1.5 and SDXL, where no VAE encoding is applied to the control signal, the Conv2d operates at the full spatial resolution (no stride, zero padding to preserve spatial dimensions).

Training masks are produced by SAM2 run on the Flux-generated training images. Training time ranges from 8h (SD1.5) to 40h (SD3.5) on two NVIDIA GeForce RTX 4090 GPUs. We will release training and inference code.

F Inference and Evaluation Protocol

Inference settings. All images are generated with the Euler sampler at 25 denoising steps, identical across SD1.5, SDXL and SD3.5 and across all baselines and DisControl. Generation resolution matches the training resolution: 1024×1024 for SDXL and SD3.5, 512×512 for SD1.5.

Training data preparation. We use $982{,}734$ images from the MONET dataset Aubin et al. (2026). Depth maps are extracted with Depth-Anything-V2-Large, Canny edge maps with OpenCV at thresholds $(100, 200)$, and surface normals with DSINE Bae & Davison (2024).

Baseline checkpoints. The vanilla scale-modulation rows in Table 4 are our own retrained ControlNets, following the official Diffusers ControlNet training examples and modeled on the most-used community ControlNet for each backbone. SmartControl and ControlNet++ use the official authors' released checkpoints with no retraining.

Evaluation metrics. Depth distances (MSE and $\min_{a,b}\mathrm{MSE}$) are computed at the native image resolution (no resizing) between the input control depth map and the depth map re-extracted from the generated image with DepthAnythingV2-Large. The Canny distance is the bidirectional Chamfer distance in pixel units between the input edge map and the edge map re-extracted from the generated image with the same OpenCV thresholds $(100, 200)$. For surface normals we report the mean angular error in degrees between the input normal map and the map re-extracted with DSINE.

Variance/mean curves. For each (backbone, modality) we evaluate 1000 images, drawing one strength per image over the whole range, and fit the mean and the spread of adherence jointly on the raw per-image measurements by heteroscedastic maximum likelihood. The two arms are paired: the same images and seeds are used for both.

CLIP score. CLIP similarity is computed with the default OpenAI `ViT-B/32` model between each generated image and its text prompt.

G Training-time region sampling

At each training step a per-pixel strength map S is drawn. With probability 0.5 a segmentation mask of an object in the image is used, after a random geometric and morphological perturbation, to split the image into the object and its complement. Otherwise, one time in five the map is constant over the whole image, and the rest of the time the image is split into two regions of a random shape (vertical

Backbone	Modality	base	ours	Δ%	win%
SD1.5	depth	4.963	6.983	**+40.7**	77
	canny	4.902	6.861	**+39.9**	79
	normals	5.083	7.064	**+39.0**	78
SDXL	depth	7.184	8.697	**+21.1**	69
	canny	5.957	8.643	**+45.1**	80
	normals	7.072	9.182	**+29.8**	77
SD3.5	depth	11.217	9.912	−11.6	31
	canny	10.642	9.626	−9.5	34
	normals	11.243	9.572	−14.9	25

Table 6: The quality result of Table 1 holds on object masks for both UNet backbones, and SD3.5 is again where it reverses. Win% is the share of the 1000 paired images on which DisControl scores higher.

		HPSv3++ ↑				CLIP ↑	
Backbone	Modality	base	ours	Δ%	win%	base	ours
SD1.5	depth	4.540	5.718	**+26.0**	68	31.021	**31.246**
	canny	3.987	5.452	**+36.7**	73	31.104	**31.265**
	normals	4.679	5.443	**+16.3**	62	31.061	**31.235**
SDXL	depth	4.605	6.117	**+32.8**	71	30.950	**31.486**
	canny	3.592	5.722	**+59.3**	80	30.839	**31.440**
	normals	4.533	6.452	**+42.3**	77	30.939	**31.346**

Table 7: The quality result is not a property of the synthetic test set: on real photographs with annotated object masks DisControl wins every cell, and CLIP similarity, which favours the baseline on the synthetic set, reverses in our favour here.

and horizontal halves, squares, circles, rectangles, ellipses and their centered variants), the first of which is also morphologically perturbed. Each region receives an independent strength value, drawn from a mixture placing 10% of its mass at $s = 0$, 20% at $s = 1$, 45% on $\mathcal{U}(0, 0.2)$ and 25% on $\mathcal{U}(0.2, 1)$. The map blends the control signal with noise pixel-wise as $S \odot c + (1 - S) \odot \eta$, and is concatenated to the degraded map along the channel dimension as an additional input to the adapter.

H Sample selection for qualitative grid figures

In Figure 2 in the main text, we show qualitative examples of the large variation in the adherence to the control map. The procedure used to select the 3 samples per strength value consists of generating 100 images and then picking the 25th, 50th, and 75th percentile images at each strength value. The strength values themselves are chosen to span a representative spectrum of adherence to the control map.

I Quality under Local Modulation: Object Masks and a Second Test Set



Figure 7: Global control modulation examples for SD1.5 and depth modality.



Figure 8: Local control modulation examples for SD1.5 and depth modality.

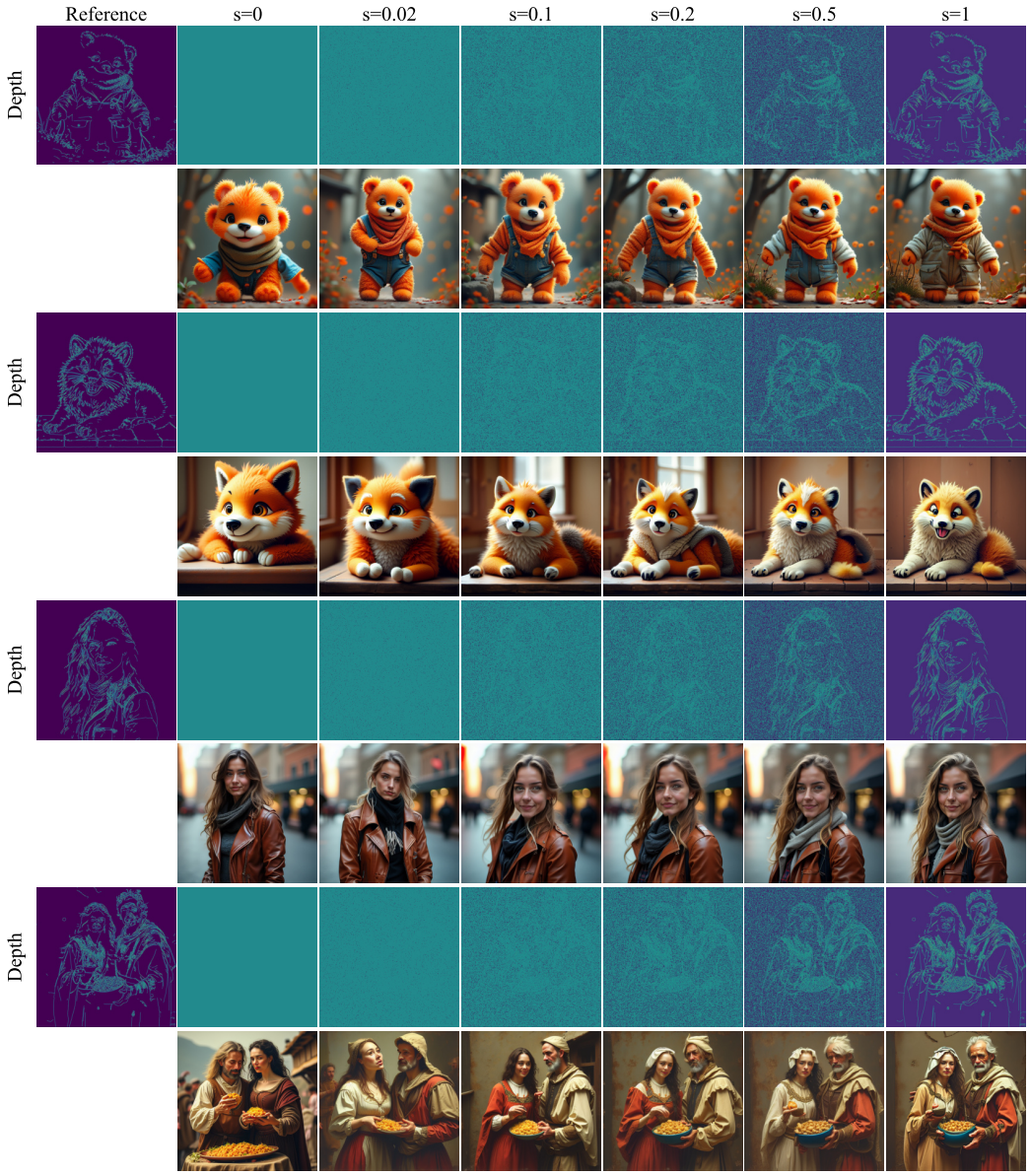

Figure 9: Global control modulation examples for SD1.5 and canny modality.



Figure 10: Local control modulation examples for SD1.5 and canny modality.



Figure 11: Global control modulation examples for SDXL and depth modality.

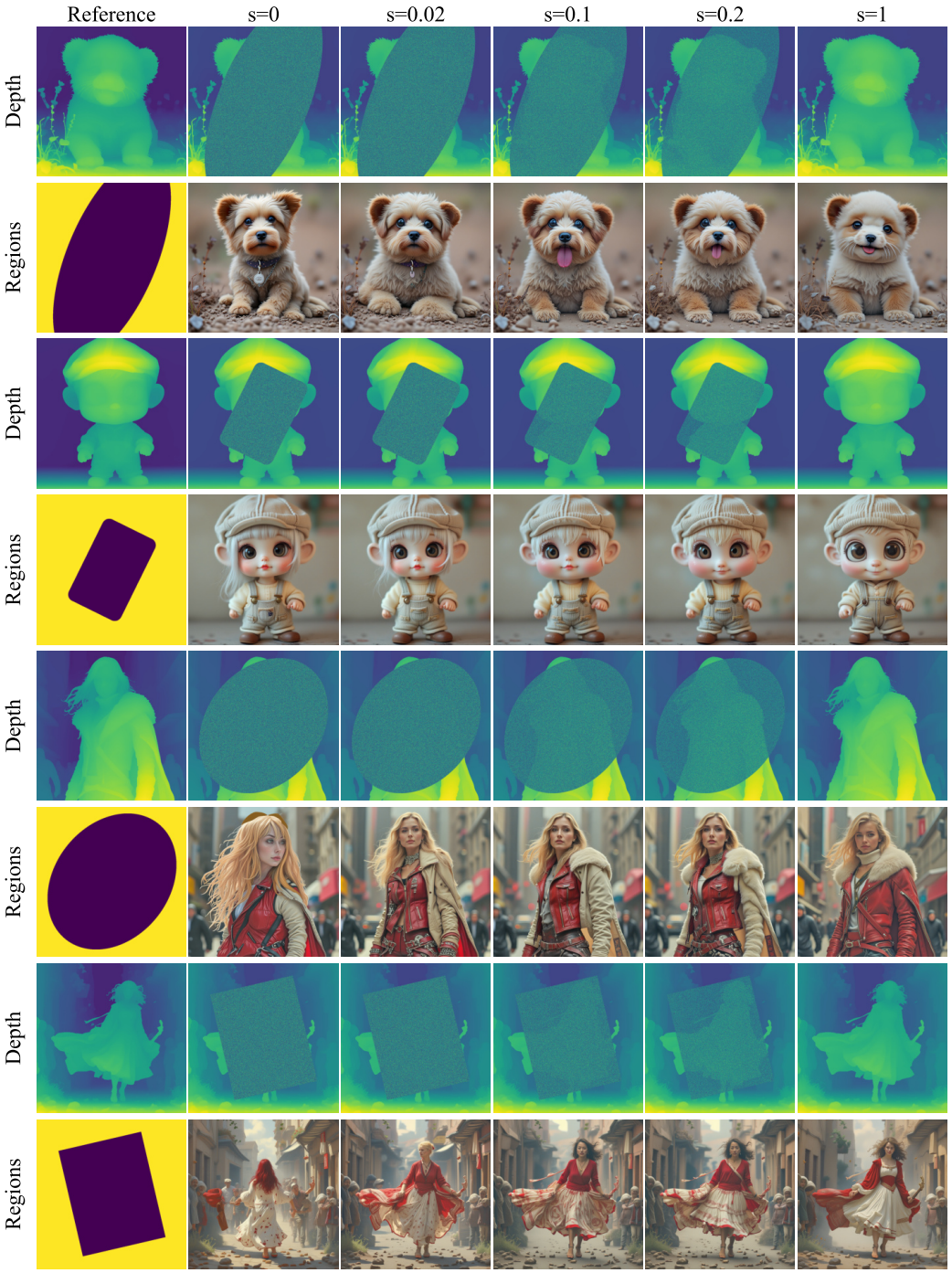

Figure 12: Local control modulation examples for SDXL and depth modality.

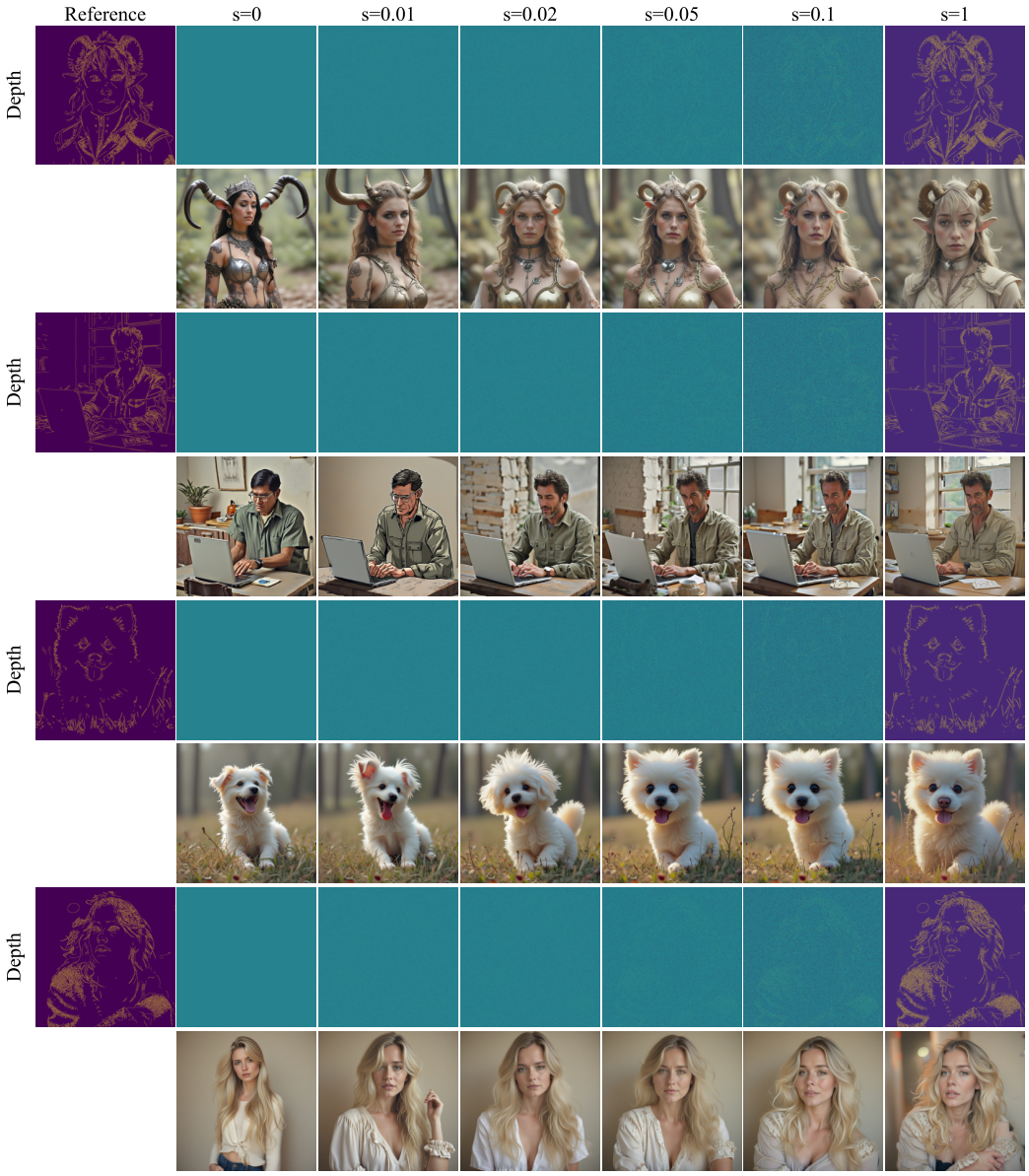

Figure 13: Global control modulation examples for SDXL and canny modality.



Figure 14: Local control modulation examples for SDXL and canny modality.

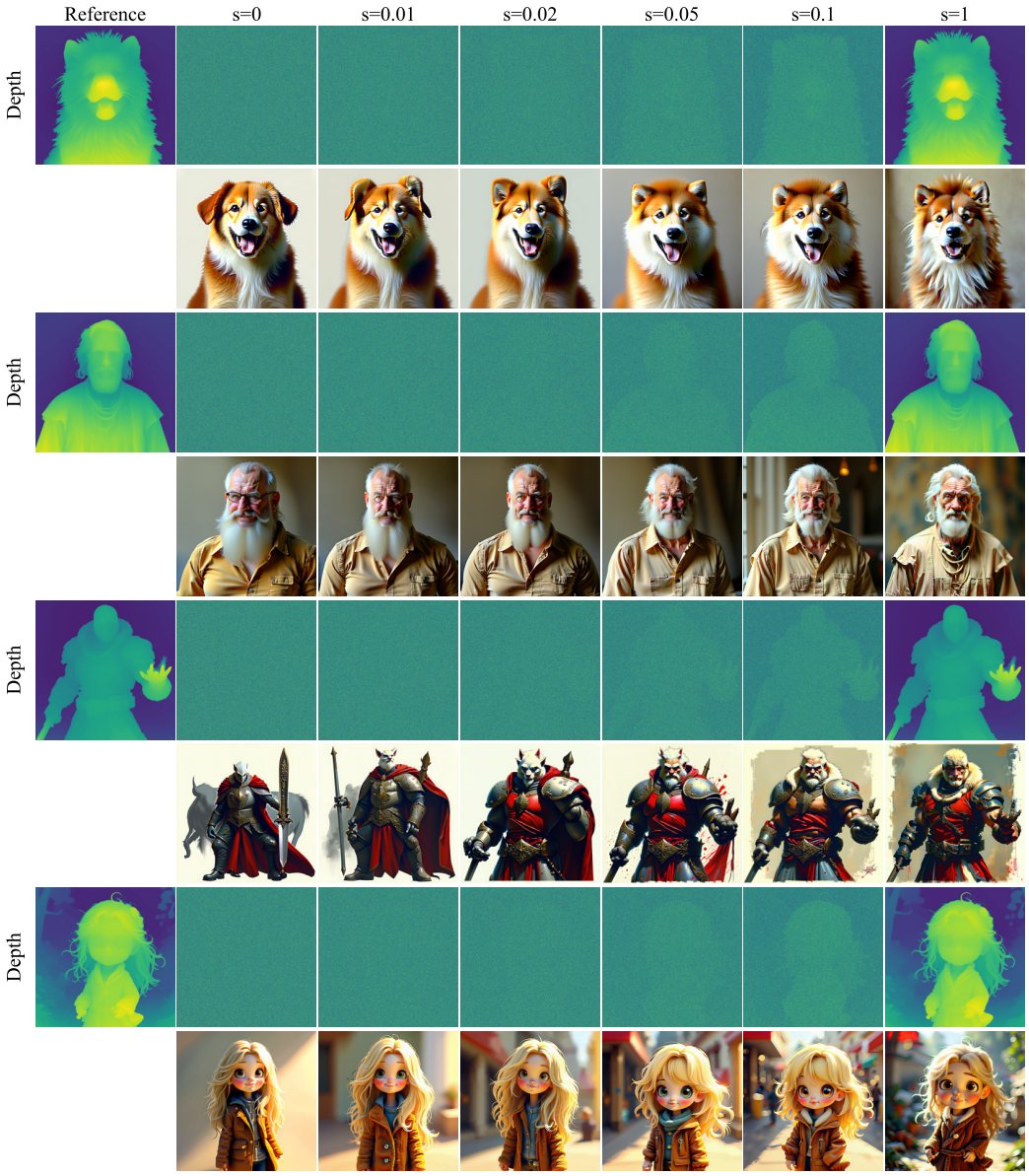

Figure 15: Global control modulation examples for SD3.5 and depth modality.

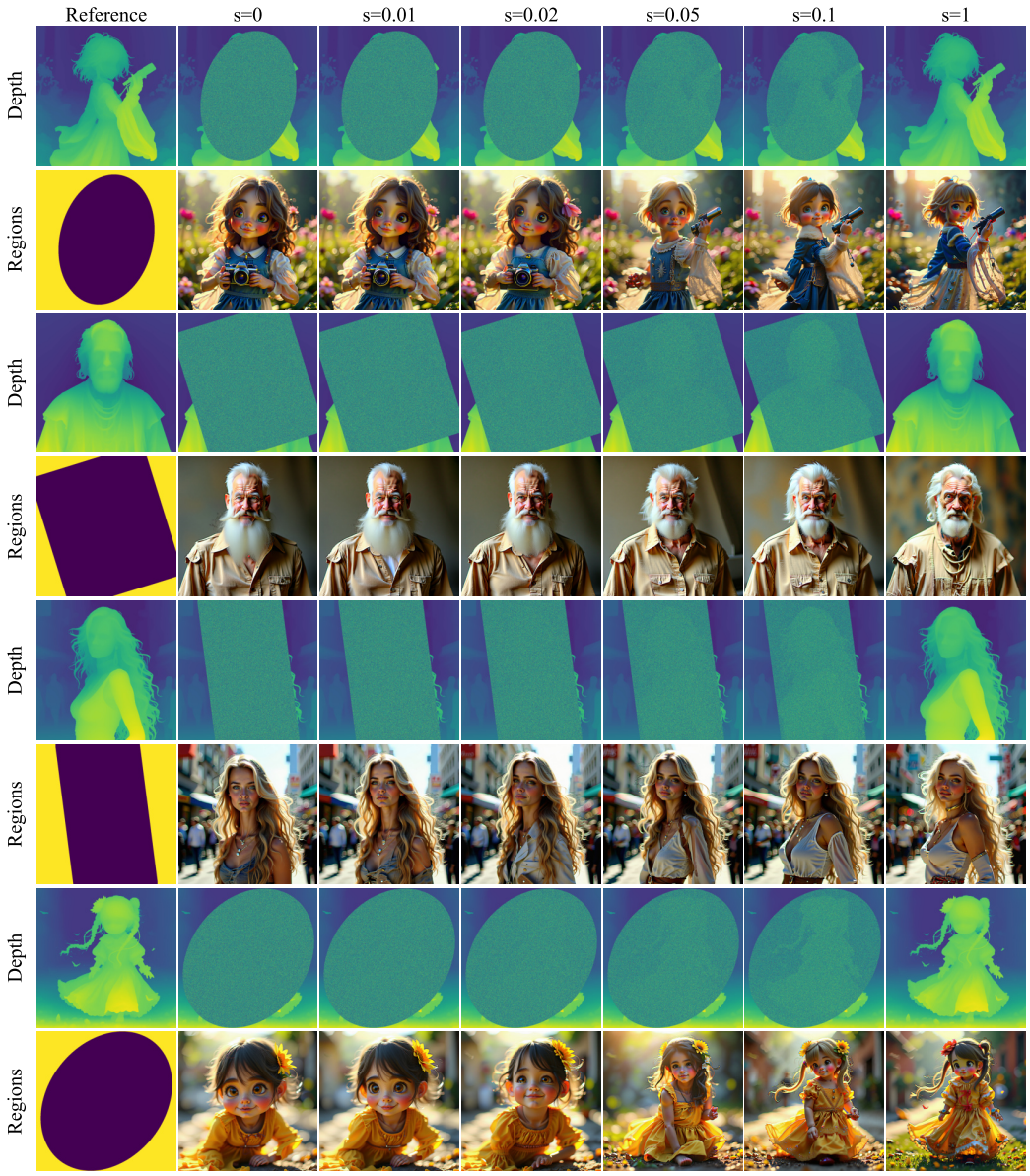

Figure 16: Local control modulation examples for SD3.5 and depth modality.

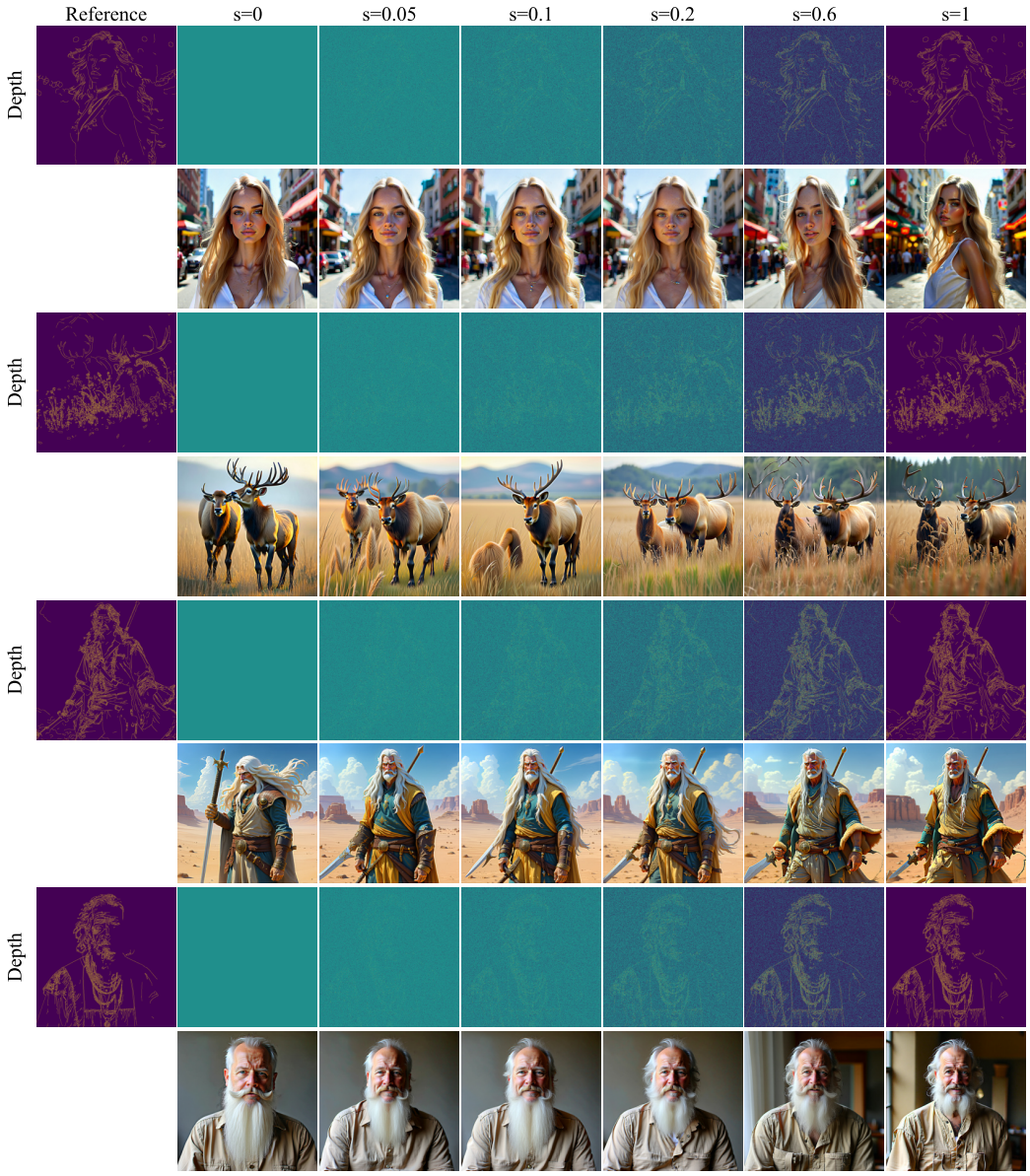

Figure 17: Global control modulation examples for SD3.5 and canny modality.



Figure 18: Local control modulation examples for SD3.5 and canny modality.

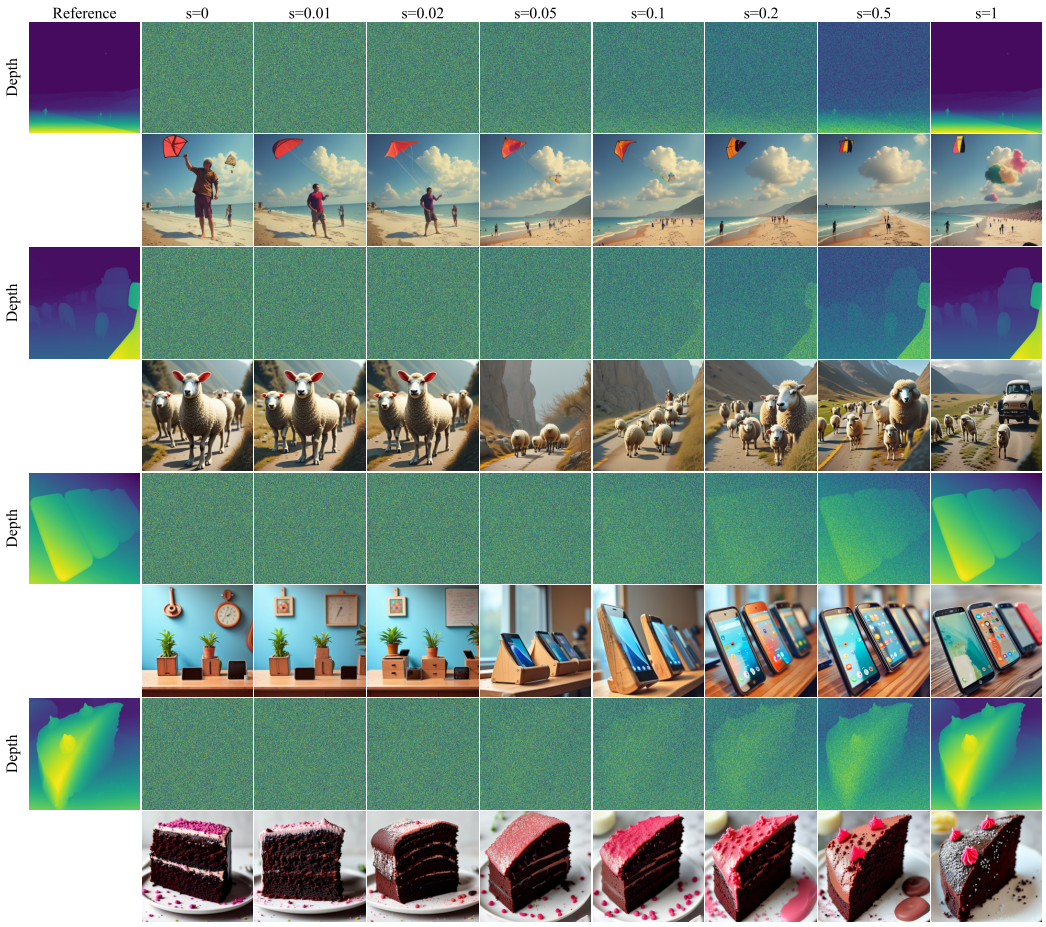

Figure 19: Global control modulation examples for SD1.5 and depth modality on the COCO dataset.

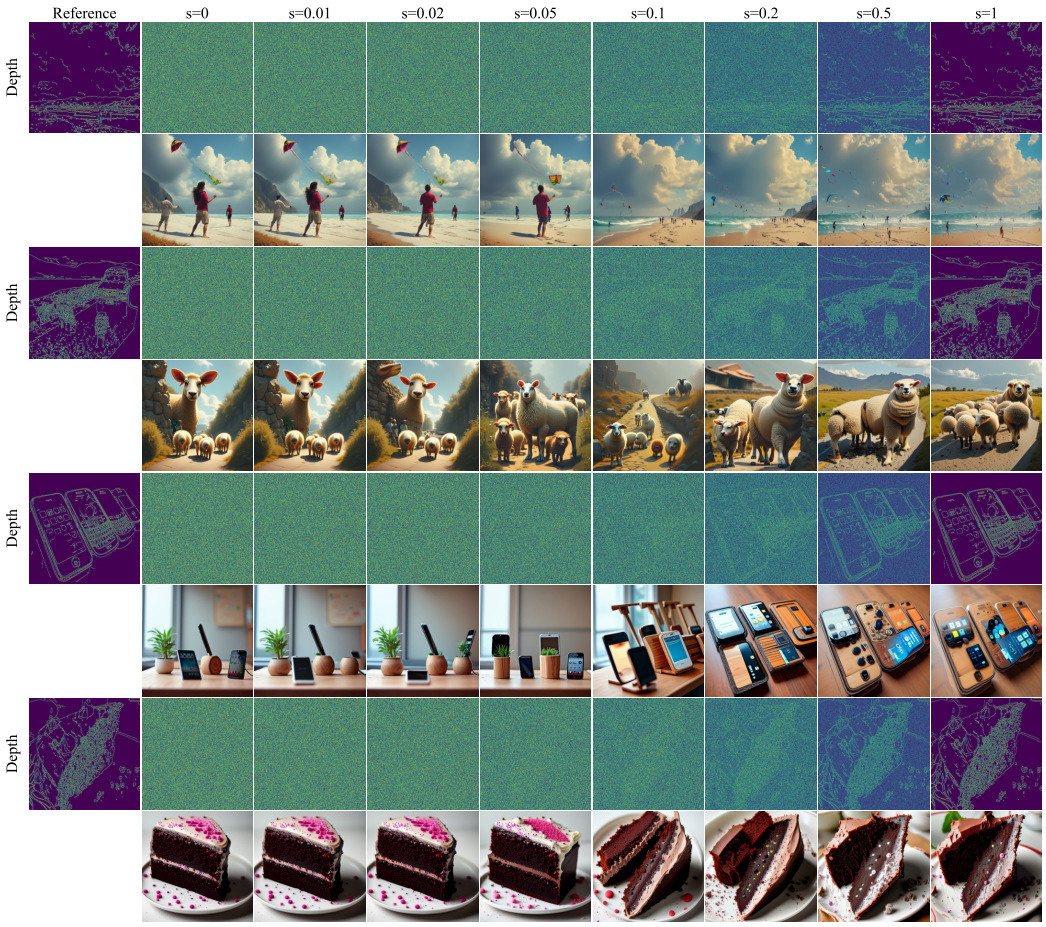

Figure 20: Global control modulation examples for SD1.5 and canny modality on the COCO dataset.

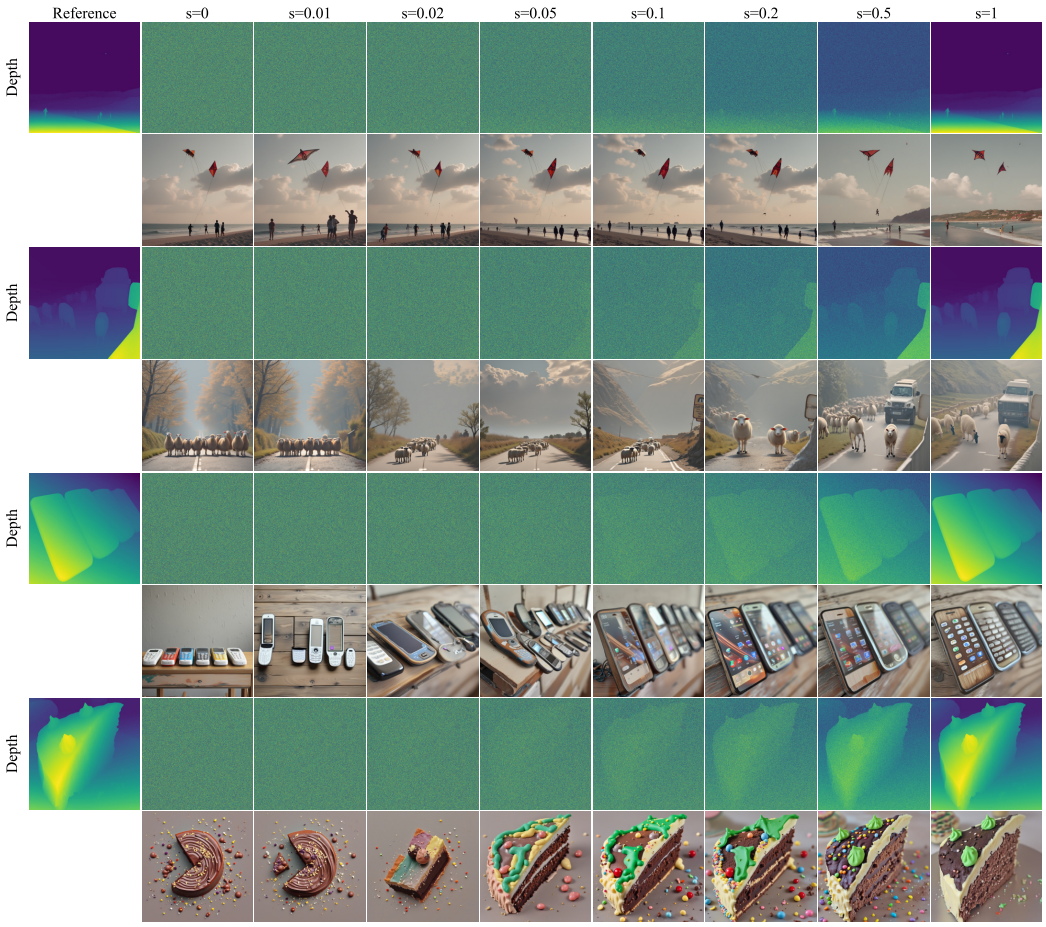

Figure 21: Global control modulation examples for SDXL and depth modality on the COCO dataset.

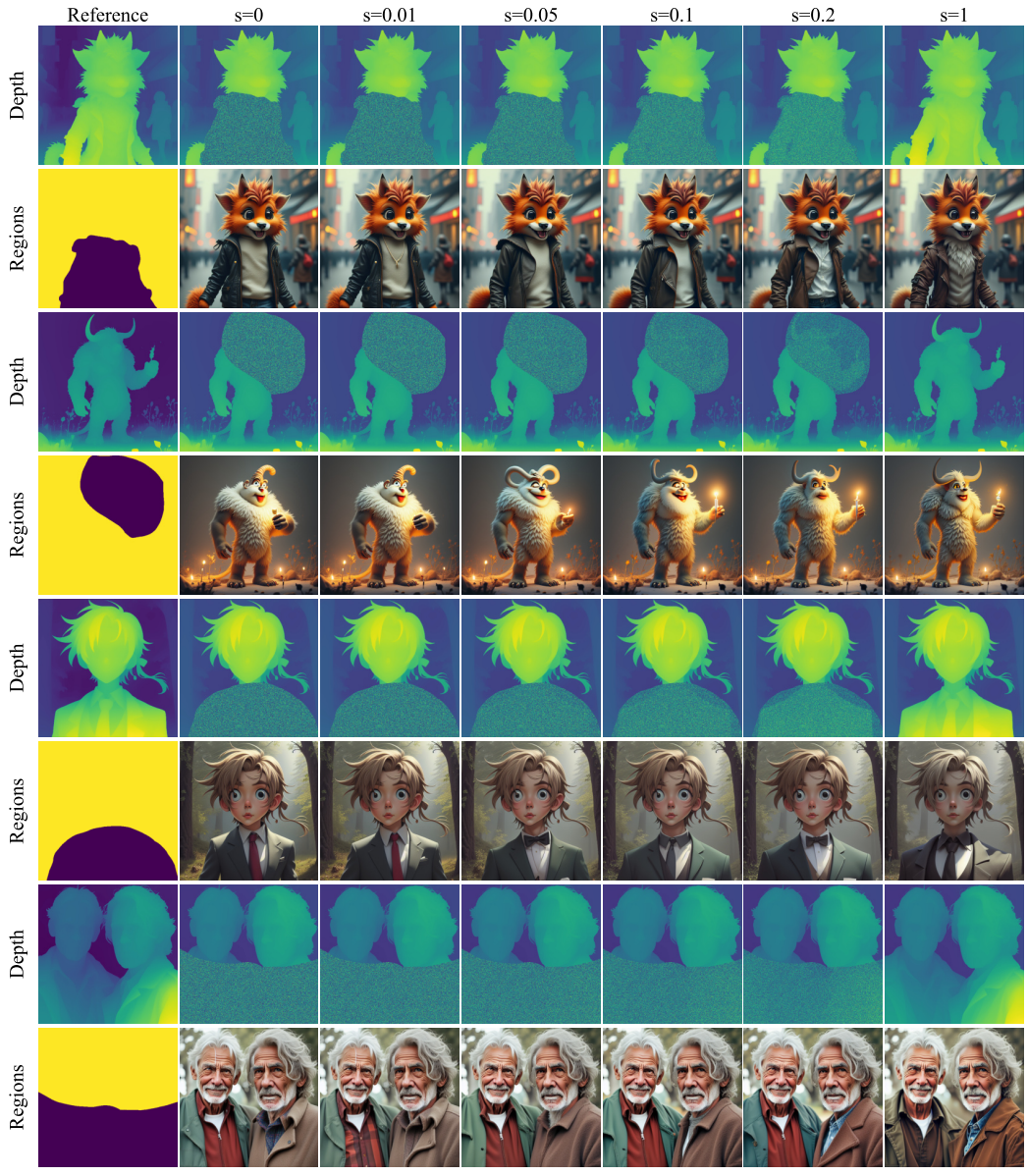

Figure 22: Local control modulation examples for SD1.5 and depth modality with hand-drawn regions.



Figure 23: Local control modulation examples for SDXL and depth modality with hand-drawn regions.

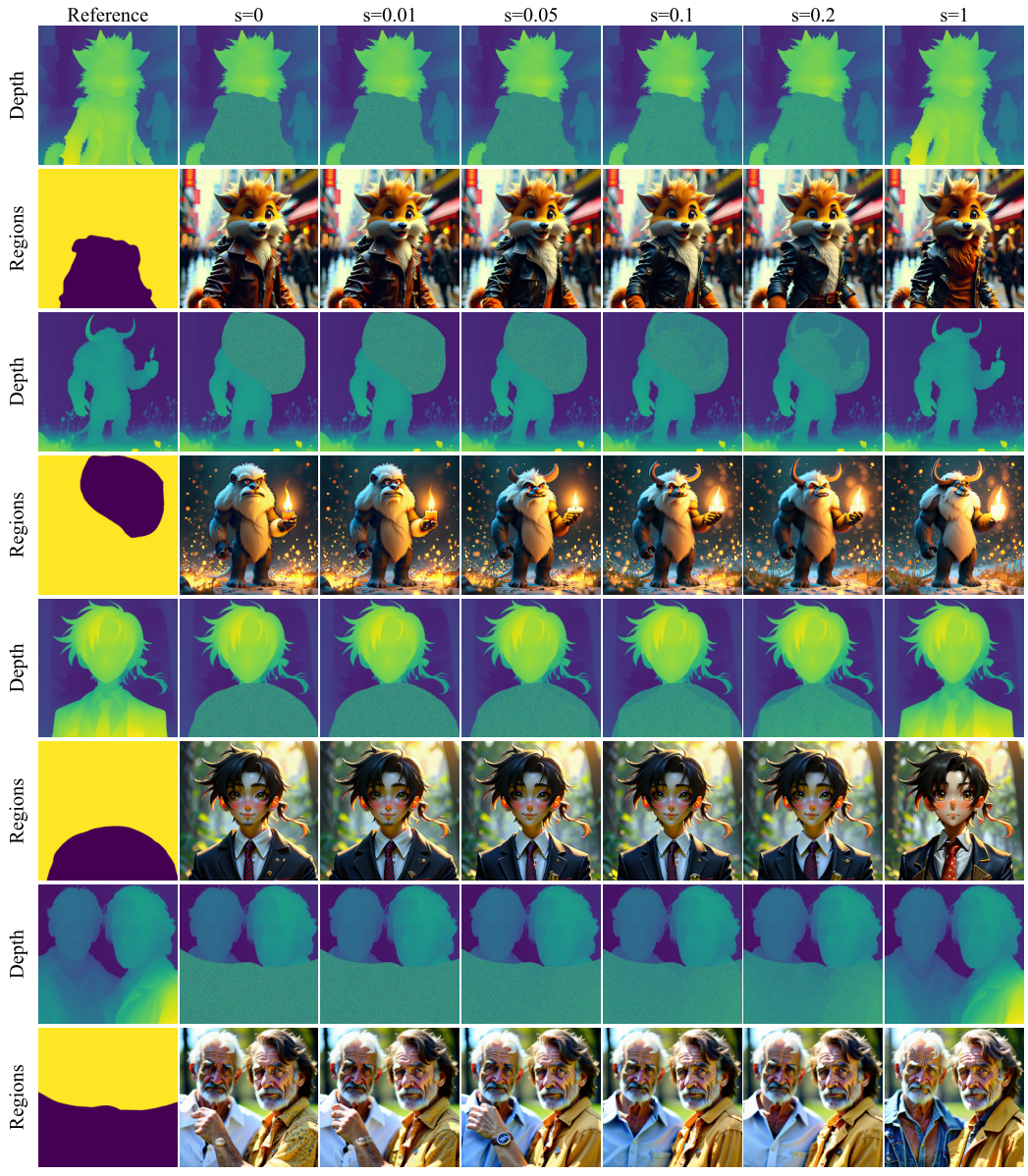

Figure 24: Local control modulation examples for SD3.5 and depth modality with hand-drawn regions.