WHERE THE EXAMPLES WERE: LOCAL LOSS WEIGHTING INHERITS THE SPATIAL LAYOUT OF SCARCE SUPERVISION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Local loss weighting emphasizes fine detail, but it applies to whatever the scarce detailed examples contain, including where they are. We test this in a controlled diffusion task where four texture classes, each in a cyan frame at one of four panel positions, share an identical low-resolution observation. Moving the same 32 high-resolution photograph–detail pairs to another panel, without adding conditions or changing the budget, moves the extra frames that region-of-interest (ROI) weighting draws at unrequested panels. Their shift toward the new panel exceeds uniform weighting's by $+81.25$, $+70.31$ and $+98.44$ percentage points in three SD1.5 training seeds, with paired intervals excluding zero; uniform weighting's own shift is $+1.56$ in every seed. The effect is confined to this unrequested structure: the same contrast on joint class-and-location success crosses zero in all three seeds. Moving only the weighted region removes the extra frames instead of relocating them, in two seeds under each of two arrangements, while the layout code moves them consistently under neither. How the same examples are arranged also changes how much the weighting helps, in an interaction whose balanced half replicates. Our readouts are frame and classifier proxies rather than human judgments of texture, the target region is marked in the coarse observation, and native 4K is a boundary test.

1 INTRODUCTION

Fine detail often occupies only a small part of an image. A coarse observation may preserve a textured surface's location while erasing its distinguishing pattern. When a generator learns these distinctions from few high-resolution examples, a whole-image loss gives the relevant region little weight. Spatially selective objectives raise that weight: wavelet-guided diffusion training emphasizes detail-rich regions (Sigillo et al., 2026), attention-guided super-resolution draws spatial cues from the low-resolution input (Moser et al., 2025), and face restoration adds losses on the eyes and mouth of aligned faces, which sit at nearly the same place in every training image (Wang et al., 2021). The added weight, however, applies to whatever the scarce examples contain, including where they are.

Consider a training set whose high-resolution textures all appear in the upper-left panel, while coarse observations and evaluation requests cover several positions. Emphasizing the texture region could strengthen the class response and, with it, an unwanted association with the upper left. The direct test is to move the same photograph–detail pairs to a different panel, without adding high-resolution conditions or changing any budget, and see what moves with them.

To separate where a detail is from what it is, we construct four texture conditions with exactly the same low-resolution observation. Each texture sits in a cyan frame that survives in the observation and reveals position, while blockwise zero-mean residuals remove texture identity. Text requests name class and position with fixed codes that the fine generator must learn. At evaluation it starts from a separate coarse generator's samples, so it is tested on new backgrounds with fixed requests.

In our main result (Figure 1), region-of-interest (ROI) weighting draws extra frames at panels the request did not ask for, and when the examples move, these frames move with them. In three independent SD1.5 training seeds, their shift toward the new position exceeds the shift under uniform weighting by $+81.25$, $+70.31$ and $+98.44$ percentage points, with paired intervals excluding zero.

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.

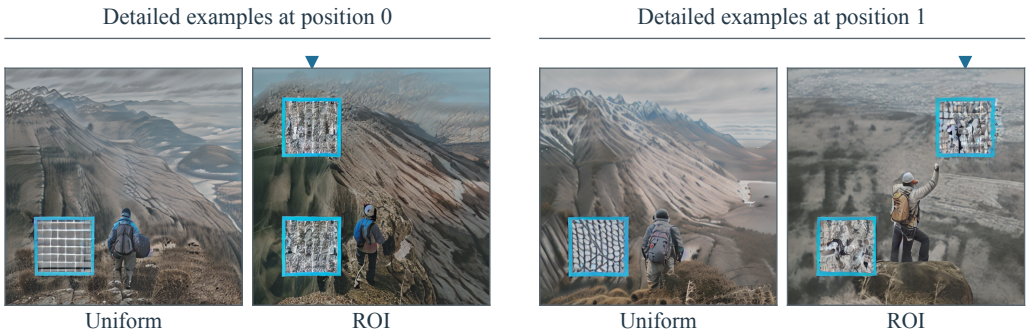

Figure 1: **The second panel appears where the detailed examples were.** One request, four models, original pixels; the requested panel is at lower left throughout. Uniform weighting draws only that panel. ROI weighting adds a second one, which moves when the training position moves (carets). Models within a group share the request, coarse condition and noise; the groups are trained separately, so their backgrounds differ. Appendix D gives the selection rule and hashes.

Under uniform weighting the shift is $+1.56$ points in every seed. A locally weighted loss thus learns where its few high-resolution examples were, to a degree that whole-image training does not; on SDXL whole-image training inherits it too, less strongly.

This inheritance has a cost that depends on how the examples are arranged. With the budget fixed at 32 photograph–detail pairs and 16,384 detailed-target visits, ROI weighting is worth more when the pairs are balanced across positions and content than when all sit at one position, in a four-model interaction positive in three seeds. The balanced gain replicates; the negative effect of concentrating the pairs does not. Changing the region coefficients partly recovers concentrated performance. Further fits move one part of the supervision at a time: moving the weighted region decides whether the extra frames survive, in two seeds under each of two supports, while moving the layout code does not consistently move them.

Our contribution is a controlled empirical finding, stated at the strength the evidence supports: *a locally weighted loss also learns where its few high-resolution examples were, and how those examples are arranged changes how much the weighting helps*. It rests on two reusable tools: a construction whose coarse observation shows where a detail is but not which one (Section 3), and a reassignment test that moves the same examples at a fixed budget (Section 5.1). We state two boundaries throughout. First, the training position carries unrequested structure, not task success; the joint-success interaction under the same contrast crosses zero in all three seeds. Second, the target region is visible in the coarse observation, so none of this shows how to locate details that are not marked. Automatic-mask controls and native 4K runs mark where the claim stops.

2 Related work

Where spatial information enters a diffusion model. Building on diffusion and latent diffusion (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024), much work organizes *inference* spatially, through fused constrained paths (Bar-Tal et al., 2023), progressive upscaling (Du et al., 2024), patch-specific prompts (Lin et al., 2025), attention guidance with pixel-space upsampling (Cao et al., 2025), or training-free extrapolation in transformers (Xue et al., 2026). Others supply spatial structure as an *input*: ControlNet conditions on edges, depth, or segmentation (Zhang et al., 2023), and GLIGEN grounds entities on boxes (Li et al., 2023). Closest to us, a third group makes the *training loss* spatially selective. Latent Wavelet Diffusion couples wavelet-guided supervision with a frequency-sensitive VAE objective (Sigillo et al., 2026), YODA steers super-resolution using low-resolution attention (Moser et al., 2025), and CTCal weights timesteps using lower-noise cross-attention (Guo et al., 2026). We intervene here, asking what a fixed regional loss does when the detailed examples it emphasizes are scarce and unevenly placed. Our high-frequency control ports one denoising-loss component with a frozen VAE; it does not reproduce the full LWD framework.

Learning incidental regularities. Discriminative models routinely latch onto regularities that hold in training data but are not the intended signal (Geirhos et al., 2020). Location is a common one: recognition degrades when familiar objects appear in unfamiliar places (Beery et al., 2018). Our setting is the generative counterpart, with a practical difference. The regularity here is the arrangement of a few permitted examples, which cannot simply be rebalanced like a dataset property, and the model reproduces it because of the loss weighting meant to improve detail.

Scarce supervision, and evaluation that separates what from where. Few-example adaptation learns a subject identifier (Ruiz et al., 2023), composes concepts under limited updates (Kumari et al., 2023), or decouples customization from preservation (Liao et al., 2026). For atypical layouts, Bao et al. (2026) disentangle semantics from primitives with a saliency-aware objective. These methods vary what is observed and which parameters move. We instead vary the spatial assignment of an unchanged set of photograph–detail pairs, which requires an evaluation that keeps content and position apart. Compositional benchmarks do this for natural prompts, covering position and count (Ghosh et al., 2023) and distinguishing attribute binding from spatial relations (Huang et al., 2023). We score class response and localization separately and then count the requests that satisfy both, on DTD-derived detail classes (Cimpoi et al., 2014) that share one low-resolution observation.

3 Isolating the role of spatial support

We vary two parts of the supervision separately: where the detailed training targets sit, and how the loss weights each target (Figure 12). The adaptation budget is fixed, so wider spatial coverage never comes with more high-resolution examples.

3.1 A detail distinction that low resolution cannot reveal

The 512×512 targets combine photographic backgrounds from the Unsplash research dataset (Unsplash, n.d.) with grayscale textures from four DTD classes: woven, knitted, meshed, and marbled (Cimpoi et al., 2014). Each texture occupies a 112×112 interior within a cyan 128×128 frame, placed at one of four positions. Text requests describe the background and name the position and detail with opaque layout and detail codes, which the generator must learn during adaptation.

For background b, layout ℓ, and detail d, write the target as $x_{b\ell d} = x^0_{b\ell} + r_{b\ell d}$. The block-averaging operator A has full row rank. The constructed residual sums exactly to zero, in integers, inside every block it intersects and is zero elsewhere, so it lies in the nullspace of A. For a uniform counterfactual choice D among the four detail classes,

$$Ax_{b\ell d} = Ax^0_{b\ell} = L_{b\ell}, \qquad I(D; L \mid b, \ell) = 0. \tag{1}$$

Figure 2 shows one such group. All four details reuse the same encoded $L_{b\ell}$, whose surviving frame locates the target without identifying the detail. Because the generator also receives a detail code, this shared observation does not cap conditional generation at chance. Appendix N proves the statement, gives the projector form, and shows that it is not a property of the VAE.

A condition specifies a photograph, a layout and a detail class; a photograph–detail pair leaves out the layout. Of the 1,024 scheduled conditions, only 32 are supervised, each giving one detail on one photograph; we call this set the support. Every other fine-training slot carries zero loss, and each use of a supervised condition counts as one high-resolution visit.

3.2 Which positions receive detailed supervision?

We compare three support designs. In a *single-position*, or concentrated, design all 32 supervised conditions sit at layout 0, or all at layout 1. The *initial spread* puts eight at each of the four layouts, but its ordering ties background content groups to layout. The *balanced spread* keeps the same 32 photograph–detail pairs and reassigns layouts to balance both content group and detail class across positions. Each layout receives eight conditions, two per detail class, and each photograph still has detailed supervision at only one position. Appendix A gives the assignment rule.

Balancing removes one measured link between content and position while the source pairs, their number and their exposure stay fixed; targets, coarse conditions, layout codes and masks change with it. Other links between position and content may remain.

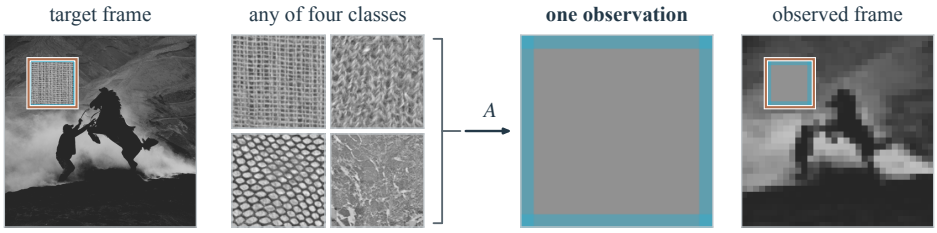

Figure 2: **Four detail classes, one observation.** The marked panel can hold any of four textures (clockwise from top left: woven, knitted, meshed, marbled), but A maps all four to the same array, so 1,024 targets share 256 observations, one per background and layout. The frame survives averaging and the texture does not, so L locates the region without identifying its content.

Reassigning the training position. The two single-position designs form a direct intervention. Moving from layout 0 to layout 1 moves four things together: the detailed targets, their coarse observations, the layout codes in the text, and the ROI derived from those observations. Everything else is unchanged: the non-detailed indices, the photograph–detail pairing, the optimizer state, the random streams, and the budget of 8,192 updates with 16,384 detailed-target visits. Every model is trained from initialization, and each pair is verified tensor by tensor before the first update and over the first 32 updates. Section 5.4 separates the four components.

3.3 How strongly is the observed region weighted?

For noisy latent z_t, caption c, and coarse latent C, define the denoising error at spatial location u, averaged over latent channels, as

$$e_\theta(u) = \frac{1}{4}\sum_{k=1}^{4}\left[\epsilon_k(u) - \epsilon_\theta(z_t, t, c, C)_k(u)\right]^2. \tag{2}$$

Uniform supervision averages this error over the latent grid Ω. ROI supervision assigns a fraction $\alpha \in [0,1]$ of the loss mass to a region M:

$$\mathcal{J}_{\mathrm{ROI}}(\alpha) = \mathbb{E}\left[\frac{\alpha}{|M|}\sum_{u\in M} e_\theta(u) + \frac{1-\alpha}{|\Omega \setminus M|}\sum_{u\notin M} e_\theta(u)\right]. \tag{3}$$

Only high-resolution slots contribute loss. With m disjoint admissible masks of area ρN, let p_j be the fraction of high-resolution visits assigned to mask $M^{(j)}$ and q_j the averaged spatial loss coefficient inside it. Writing $c_{\mathrm{out}} = (1-\alpha)/(1-\rho)$ and $\kappa = (\alpha-\rho)/[\rho(1-\rho)]$,

$$\mathbf{q}_S - \left(c_{\mathrm{out}} + \tfrac{\kappa}{m}\right)\mathbf{1} = \kappa\left(\mathbf{p}_S - \tfrac{1}{m}\mathbf{1}\right). \tag{4}$$

At $\alpha = \rho$ the weighting is exactly uniform. For $\alpha > \rho$, a position that holds more of the detailed examples receives proportionally more loss weight, with positive factor κ. The reference configuration uses $\rho = 1/16$, $\alpha = 1/2$, and $m = 4$. Initial and balanced spread have the same average spatial weights but pair masks with different targets and denoising errors. Appendix N gives the derivation and the coefficients.

To locate M, we decode L and choose the best-matching frame among the four public geometries by frame-pixel error, without inspecting the texture label. This cue is task-specific, and the weighted region contains both frame and texture. We therefore also test separate texture-interior and frame-ring coefficients, $\alpha \in \{1/4, 1/2, 3/4\}$, and centered 12^2, 16^2 or 20^2 masks, in twelve fits with seed 0 and both supports (Appendix I). Changing mask area also changes coverage and per-pixel weight, so area is not isolated.

3.4 A COMMON GENERATION AND OPTIMIZATION PROTOCOL

The fine generator is an SD1.5 latent-diffusion U-Net with rank-8 LoRA (Hu et al., 2022) and a trainable coarse-conditioning input branch. It is trained on the observed coarse latent and evaluated on generated coarse samples from a frozen bank shared by the compared models.

Each main fit receives 8,192 optimizer updates with a scheduled batch of 16, and 16,384 high-resolution visits, 512 per condition. The remaining 114,688 slots keep their place in the schedule and the random sequence but contribute zero loss. Within each seed, uniform and ROI fits share initialization, data order and every random draw, as well as model size and trainable parameter count. Appendix A gives AdamW (Loshchilov & Hutter, 2019) and the remaining settings.

Comparison of interest. With J the joint class-and-location success, support S, weighting w and fine seed r, we estimate $\Delta(S, w, r) = J_w(S, r) - J_{\text{uniform}}(S, r)$ and

$$I(w, r) = \Delta(S_{\text{balanced}}, w, r) - \Delta(S_{\text{single0}}, w, r), \tag{5}$$

so I measures the support–weighting interaction directly. Its interval resamples the same photographs for all four models, and seeds are reported separately.

4 MEASURING CLASS RESPONSE AND SPATIAL CONTROL

We evaluate each 512^2 model on 2,048 fixed requests: 64 held-out backgrounds, four layouts, four details, and two noise repetitions. Four detail requests with the same background, layout, and noise form a quartet, giving 512 quartets. Sampling uses 30 DDIM steps (Song et al., 2021) and guidance 7.5, with the same coarse samples and initial noise for all models in a comparison.

Four complementary scores. A frozen VGG-descriptor/SVM classifier (Simonyan & Zisserman, 2015) scores class response, and the fixed frame detector scores localization. Let y_{gd} indicate correct class and location for detail d in quartet g, counting invalid outputs as failures. With $G = 512$,

$$J = \frac{1}{4G}\sum_{g=1}^{G}\sum_{d=1}^{4} y_{gd}, \qquad A_4 = \frac{1}{G}\sum_{g=1}^{G}\prod_{d=1}^{4} y_{gd}, \tag{6}$$
$$\max(0, 4J - 3) \le A_4 \le J.$$

J is the fraction of individual requests that succeed, and A_4 (ALL4) is the fraction of quartets in which all four requests succeed. The bounds need no independence assumption (Appendix N). Class response does not certify texture fidelity, and frame detection does not measure the texture's full spatial extent.

A frozen classifier checked on held-out textures. The classifier was trained, selected and confirmed in three stages. It was trained on 158 development photographs, chosen between two candidates by a fixed rule on 156 validation photographs (126 correct), and, once frozen, tested on 60 separate confirmation examples: 45 correct, with class counts $9, 12, 12, 12$ out of 15. This meets the prespecified threshold of 45 exactly and exceeds the classwise minimum of 8. No input was invalid, and the classifier was not refit.

We then rescored 26 completed generator models with it, on a list fixed before the confirmation outcome was known and with sample order, frame decisions, invalid handling and input and descriptor hashes unchanged. Two earlier classifiers had failed confirmation, at 23/32 and 42/64 (Appendix C), and the twelve later fits are scored with the same frozen replacement. Confirmation covers constructed grayscale textures, not generated outputs or images without texture, so classifier errors that depend on the training scheme remain possible even with the readout frozen. A blinded human panel run as a check did not reach the classifier's bar. Three untrained raters scored 64.6% where the classifier confirms at 75%, and agreed with each other on 66.0% of real textures but only 24.3% of generated crops, so our readouts remain classifier- and geometry-based (Appendix M).

Paired uncertainty. Reported 95% intervals use 2,000 percentile-bootstrap draws over photographs, keeping all requests from each sampled photograph. They describe variation across held-out photographs for fixed models and a fixed readout. Training-seed variation and classifier error lie outside

Table 1: **Where extra frames sit follows the training position in every seed; the per-request readouts do not.** Counts out of the same 64 requests, before and after the detailed examples move from position 0 to 1, under the rule of Appendix H. K is the shift of extra frames toward position 1 under ROI weighting minus the same shift under uniform weighting. Joint success crosses zero in all three seeds, the other two in two of three.

		Extra frame at		Per request		
Seed	Loss	position 0	position 1	Requested frame	Any extra frame	Joint success
0	Uniform	$2 \to 1$	$2 \to 2$	$47 \to 55$	$5 \to 5$	$14 \to 21$
	ROI	$30 \to 1$	$2 \to 26$	$25 \to 37$	$31 \to 28$	$3 \to 13$
	K **(pp)**	$+\mathbf{81.2}$ $[+64, +98]$		$+6.2$ $[-9, +22]$	-4.7 $[-22, +12]$	$+4.7$ $[-11, +20]$
1	Uniform	$1 \to 1$	$2 \to 3$	$55 \to 50$	$5 \to 6$	$20 \to 16$
	ROI	$25 \to 2$	$2 \to 25$	$42 \to 39$	$29 \to 26$	$9 \to 10$
	K **(pp)**	$+\mathbf{70.3}$ $[+53, +88]$		$+3.1$ $[-12, +19]$	-6.2 $[-23, +11]$	$+7.8$ $[-9, +25]$
2	Uniform	$1 \to 1$	$2 \to 3$	$54 \to 52$	$5 \to 5$	$17 \to 19$
	ROI	$27 \to 1$	$3 \to 41$	$36 \to 44$	$30 \to 42$	$10 \to 13$
	K **(pp)**	$+\mathbf{98.4}$ $[+78, +116]$		$+15.6$ $[+2, +30]$	$+18.8$ $[+5, +33]$	$+1.6$ $[-17, +20]$

them, which is why seeds are shown separately and never pooled. Appendix H states the audit's selection rule, thresholds and descriptors in full.

5 What local weighting inherits from its supervision

This section moves the emphasized examples, and then intervenes on the weights themselves, to find what a regional loss carries over from those examples. Section 6 then fixes the position and asks how much the weighting helps.

5.1 Reassigning the training position relocates the extra frames

We move the high-resolution examples from the upper-left to the upper-right panel, adding no conditions and changing no budget (Section 3.2). If local weighting only sharpened the requested texture, the move would leave the rest of the image unchanged.

It does not. Table 1 counts extra frames at the two upper positions for the 64 requests placed in the two lower ones, so both upper positions are unrequested throughout. Under ROI weighting the extra frames follow the training position: in seed 0 they go from $30/64$ to $1/64$ at position 0 and from $2/64$ to $26/64$ at position 1, and the two further seeds show the same reversal. Under uniform weighting the counts stay between $1/64$ and $3/64$. The interaction K is the shift of extra frames toward the new position under ROI weighting minus the same shift under uniform weighting, on a scale from -400 to 400. It is $+81.25$, $+70.31$ and $+98.44$ points in the three seeds, and $+67.97$, $+59.38$ and $+75.00$ over all 128 requests; all six intervals exclude zero. The plan and stopping rule were fixed before the two further seeds were trained (Appendix H). The relocation grows with the weighting: K is $+32.81$ $[20.31, 46.88]$ at $\alpha = 1/4$ and $+120.31$ $[103.12, 137.50]$ at $\alpha = 3/4$ (Appendix L). On SDXL at 512^2, in one seed, K is $+39.06$ $[23.44, 54.69]$, but uniform weighting there moves the extra frames too ($M = +139.06$), and the joint-success interaction again crosses zero (Appendix K). Figure 3 shows the pixels behind the counts (more in Appendix B).

The relocation also appears without counting frames. The audit records six frame-free descriptors of the pixels at each position, and K computed on them reproduces the relocation for four: contrast, gradient energy, high-frequency energy and high-frequency fraction shift the same way, with intervals excluding zero in all three seeds. Directional anisotropy and spectral entropy do not. This check removes the dependence on the detector, but the frame's own pixels still enter the descriptors.

Where the extra frames sit follows the training position; the outcome of the request does not. The last two columns of Table 1 read the same models two other ways: joint success crosses zero in every seed, and producing the requested frame does so in two of three. All 2,048 requests agree. In the two further seeds the joint contrast is -7.03 $[-10.94, -3.08]$ and -0.54 $[-4.88, 3.76]$, and the two

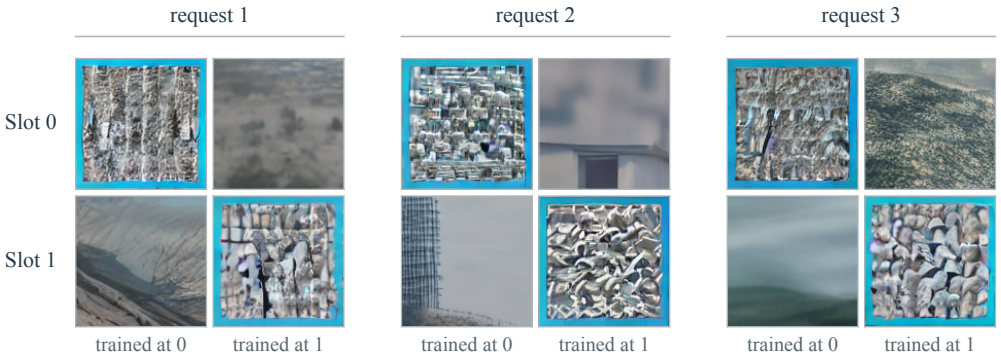

Figure 3: **What the weighted models put in the two unrequested slots.** The two upper panel regions at native 128×128 pixels, for the ROI models trained at position 0 and at position 1. The framed panel sits on the diagonal: slot 0 when the detailed examples did, slot 1 when they moved. Neither slot was requested. Requests are the first three of the six meeting the criterion of Appendix D.

values come from different metrics moving in opposite directions. Frame detection describes frame pixels, not texture (Section 4).

5.2 The concentrated-support loss is in localization, not class response

The same asymmetry appears without moving anything. With support concentrated at position 0, reference ROI barely changes class response (36.62% to 37.55%) but cuts detected localization from 83.20% to 51.51% ($[-36.72, -26.22]$), and support at position 1 splits the same way (Appendix B). For class and location events E_c, E_ℓ, let $W_a = \Pr_a(E_c \cap E_\ell^c)$ be the rate of the right class in the wrong place. Joint success then decomposes as $J_a = \Pr_a(E_c) - W_a$, so $\Delta J = \Delta C_{\text{cls}} - \Delta W$. For position-0 support, W rises from 4.98% to 13.82%, a $+8.84$-point loss that outweighs the $+0.93$-point class gain. These identities decompose the score; they do not show that texture is displaced. Figure 1 shows the coexistence they quantify: the requested panel is produced, and a second appears beside it.

Without the requirement that the requested frame beat competing positions, the interaction stays positive in two of three seeds, at $+17.09$ ($[13.04, 21.14]$) in seed 0, but not in seed 2. Neither rule validates texture (Appendices I and J).

5.3 Reallocating the region weight does not remove the extra frames

The reference mask weights both the texture and its visible frame, which points to the frame term as the cause. Two coefficient interventions on unchanged targets test this, with background weights adjusted so that the mean loss weight stays one. Weighting the texture interior by eight and the frame ring by one helps: concentrated joint success rises to 39.50%, $+15.77$ points over reference ROI ($[12.50, 19.04]$), and localization recovers to 69.53%, though it stays below uniform's 83.20%. The reverse assignment leaves a joint effect crossing zero, so a heavy frame coefficient alone does not reproduce the collapse (Table 3). Moving the extra weight inward at fixed total weight gives four-model contrasts of -4.44 $[-8.89, 0.10]$ and $+3.66$ $[-0.49, 7.86]$, opposite in direction and crossing zero.

The sharpest version of this test removes the frame from the weighted region altogether. Holding the texture coefficient at its reference value of eight, we give the frame ring exactly the background coefficient that fixed total weight determines, $632/975$, and train all four position arms from initialization in two seeds. The frame stays in the target and in the observation, so the region remains locatable. The inheritance survives: K is $+73.44$ $[56.25, 90.62]$ and $+39.06$ $[23.44, 56.25]$, against $+70.31$ and $+98.44$ when the frame is weighted. In the first seed the extra frames still go from $25/64$ to $1/64$ at position 0 and from $2/64$ to $26/64$ at position 1. Frame weighting is therefore not needed for the relocation, although the effect is smaller in the second seed. The frame remains the locator and the counted object, and the background coefficient necessarily rises.

5.4 WHICH PART OF THE MOVED SUPERVISION CARRIES IT

Reassignment moves four things together: the target, its observation, the layout code in the text and the mask derived from the observation. Further pairs of fits move one and hold the rest, against the same position-0 reference. All four are properties of one training row, so these fits are cue conflicts, not clean isolations.

Under concentrated support moving the layout code does little: K is -7.81 $[-21.88, 4.69]$ and $+12.50$ $[1.56, 25.00]$, the counts barely move, and all six frame-free descriptors cross zero in both seeds. Holding the code fixed while the rest moves reproduces the effect in full, $K = +81.25$ $[60.94, 101.56]$ and $+78.12$ $[59.38, 96.88]$, and the two parts add up: $K(\text{code}) + K(\text{rest}) - K(\text{full})$ crosses zero in both seeds and strata. Moving only the mask differs in kind. K is $+37.50$ in both seeds, but the counts are $1/3$ and $1/2$: position 0 falls from 25 and 27 to 1, while position 1 gains almost nothing. Because K compares the two positions, a positive value can come from either end, and here it comes from suppression; K on any extra frame is -40.62 in both seeds.

We repeat both cells in two seeds with the eight-photograph support of Section 6.1, where the weighting helps rather than hurts. The mask result holds and the code result does not. Moving the mask alone returns the counts exactly to the uniform model's in both seeds, $1/2$ against $9/4$, with K on any extra frame at -12.50 $[-21.88, -3.12]$ each time. It also removes the gain: ALL4 falls from 175 and 126 to 0 of 512. Moving the layout code alone raises extra frames at the trained position from 9 to 24 in one seed ($K = +23.44$ $[12.50, 35.94]$) but only to 11 in the other ($+3.12$ $[-9.38, 15.62]$). In both seeds it costs part of the gain, with ALL4 falling to 116 and 73, and only under ROI weighting (interaction -10.35 $[-14.85, -5.47]$). In all four fits, moving the mask decides whether the extra frames survive but not where they appear, and moving the layout code moves them consistently under neither support.

6 WHAT THE ORGANIZATION OF THAT SUPERVISION IS WORTH

Inheriting a spatial arrangement is not by itself a cost. Holding the training position fixed, we compare uniform and ROI weighting under two arrangements of the same supervision, then vary the weighting.

Table 2: **Spatial support changes the reference ROI effect.** All six pairs use 32 high-resolution conditions and 16,384 visits at 512^2, with $\alpha = 1/2$ and $\rho = 1/16$, evaluated on 2,048 requests. Differences are ROI minus uniform, with paired photograph-bootstrap intervals; ALL4 counts complete quartets out of 512. Seeds are fine-training replicates and are never pooled.

		Joint success			Diagnostics (pp)		ALL4
H support	Seed	Unif. (%)	ROI (%)	Δ (pp), 95% CI	Δ class	Δ loc.	Unif. → ROI
Single position 0	0	31.64	23.73	$\mathbf{-7.91}$ $[-10.99, -4.44]$	$+0.93$	-31.69	$2 \to 8$
Single position 1	0	38.57	32.81	$\mathbf{-5.76}$ $[-9.72, -1.71]$	$+4.44$	-25.98	$5 \to 9$
Initial spread	0	40.38	45.70	$\mathbf{+5.32}$ $[2.05, 8.64]$	$+7.08$	-3.17	$4 \to 5$
	1	31.15	50.20	$\mathbf{+19.04}$ $[15.87, 22.27]$	$+21.63$	-3.61	$0 \to 25$
Balanced spread	0	27.10	39.89	$\mathbf{+12.79}$ $[9.72, 15.72]$	$+13.96$	-0.68	$1 \to 8$
	1	30.13	45.95	$\mathbf{+15.82}$ $[12.94, 18.51]$	$+18.21$	-1.56	$1 \to 15$

6.1 SUPPORT ORGANIZATION CHANGES THE VALUE OF ROI WEIGHTING

With half the loss mass on the framed region, ROI weighting reduces joint success by 7.91 or 5.76 points when the supervision is concentrated at position 0 or 1, and improves it in both initial-spread and both balanced-spread fits. All six paired intervals exclude zero (Table 2; absolute levels in Appendix B).

The balanced-versus-concentrated interaction is $+20.70$ points ($[16.21, 25.00]$) for seed 0 and stays positive at $+12.11$ and $+5.91$ in two further seeds, whose twelve arms are in Table 12. Only its balanced half replicates: in the two further seeds the concentrated effect is $+3.71$ and $+2.25$ rather than negative. We therefore report the interaction and treat the seed-0 sign reversal as a single-seed observation (Appendix J).

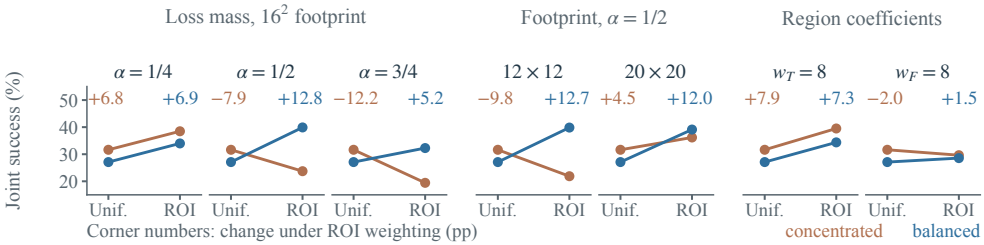

Figure 4: **Balanced support benefits from every weighting we tested; concentrated support does not.** Joint success under uniform and ROI weighting for the seven configurations, shared scale, seed 0. The interaction is the gap between the slopes at ROI; Table 9 gives every interval.

Arranged differently, the same 32 conditions make ROI weighting far more effective, and the inheritance remains. Using four details on each of eight photographs, rather than one on each of thirty-two, turns ROI weighting from harmful to decisive in two seeds: ALL4 rises from 0 to 175 and 126 of 512, and joint success from about 24% to 68.75% and 59.47%. The weighting still adds extra frames at the training position, by $+12.50$ points in both seeds. Joint success on those requests still rises, by $+37.50$ and $+21.88$, even as the requested frame falls (-6.25 crossing zero, -17.19 excluding it). Moving this support to position 1 shifts the extra frames in both seeds ($K = +34.38$ $[21.88, 48.44]$ and $+12.50$ $[3.12, 23.44]$), while ROI weighting still lifts ALL4 to 110 and 59 of 512. The frames leave the old position in both seeds but reach the new one only in the first: the counts at positions 0 and 1 go from $9/4$ to $1/19$ there, and to $2/5$ in the second. The inheritance thus comes with the benefit.

6.2 Which weighting configurations help depends on support

Concentrated support can still benefit from ROI weighting (Figure 4). At $\alpha = 1/4$ both supports improve by about seven points, whereas $\alpha = 3/4$ takes the concentrated effect to -12.21 and leaves balanced support at $+5.18$. At fixed $\alpha = 1/2$, balanced effects stay positive for 12^2, 16^2 and 20^2 masks while enlarging the mask turns the concentrated effect positive (Appendix I). None of the automatic masks we tested recovers the gains of the visible frame (Appendix B); without the frame, every fit, uniform or ROI, draws the texture at its training position (Appendix L).

6.3 Native 4K as a boundary test

Native 4096×4096 runs use SD1.5 and SDXL with constructed directional patterns and backbone-specific detectors. The one complete SDXL four-model set (seed 1, 8,192 updates from initialization) is listed in Table 4 above earlier 1,024-update continuations, whose smaller counts reflect that budget. In the complete set, ROI weighting raises joint success under *both* supports, so the support interaction is -3.65 $[-8.85, 1.56]$ on the matched positions and -2.73 $[-7.42, 2.34]$ on all requests. It crosses zero and runs opposite to the 512^2 result, so the interaction of Section 6.1 does not replicate here.

The side effect of Section 5 does appear at 4K. Detected detail in unrequested regions rises under ROI weighting in both supports, by $+13.54$ $[6.77, 20.31]$ and $+8.85$ $[3.65, 14.06]$. Forced choice does not establish detail presence, no model solves an ALL4 quartet, and the reassignment was not repeated at 4K, so this shows that the side effect is consistent, not that it replicates.

7 Conclusion

A locally weighted loss learns where the few high-resolution examples it emphasizes were, and how those examples are arranged changes how much it helps. Moving the examples moves the extra frames the weighting draws, by 70 to 98 points of interaction in three SD1.5 seeds where whole-image training moves none, and in two seeds where the weighting helps most. The weighted region decides whether the extra frames survive; the layout code does not consistently move them. Two checks follow for anyone adopting a local loss: move the scarce examples to see whether side effects follow, and balance them across positions and content, where every weighting we tested helped.

REPRODUCIBILITY STATEMENT

The construction, support interventions, optimizer schedule, sampling requests, and readout history are described in Sections 3–4 and Appendices A and I. Tables and quantitative figures are generated from versioned result summaries. The experiment archive records dataset and model identifiers, paired random streams, checkpoint ancestry, execution status, and evaluation records. Anonymized code for all experiments is provided as supplementary material; data, pretrained models, checkpoints and generated images are not included.

ETHICS STATEMENT

The work studies a controlled adaptation and measurement problem using existing image sources. Five volunteer raters, all of whom gave informed consent, labelled texture crops for a blinded check of the class readout (Section 4); the task involved no sensitive content, and only aggregate results are reported. Generated details should not be presented as factual content recovered from the coarse observation. Any artifact release must respect the original terms for source photographs and pretrained models and distinguish construction metadata from image files with redistribution restrictions.

AI USE STATEMENT

The authors used generative AI tools to assist with manuscript drafting and polishing, literature review, experimental planning and implementation, and the analysis and visualization of recorded experimental results. Some source captions contain machine-generated descriptions. The authors take full responsibility for the research decisions, reported results, and final manuscript.

REFERENCES

Nan Bao, Yifan Zhao, Wenzhuang Wang, and Jia Li. Envisioning beyond the few: Disentangled semantics and primitives for few-shot atypical layout-to-image generation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.31266.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/bar-tal23a.html.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision*, pp. 472–489, 2018. doi: 10.1007/978-3-030-01270-0_28. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Beery_Recognition_in_Terra_ECCV_2018_paper.html.

Boyuan Cao, Jiaxin Ye, Yujie Wei, and Hongming Shan. RepLDM: Reprogramming pretrained latent diffusion models for high-quality, high-efficiency, high-resolution image generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/58a16bd2eb860f119a07aa98195c7672-Abstract-Conference.html.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Cimpoi_Describing_Textures_in_2014_CVPR_paper.html.

Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. DemoFusion: Democratising high-resolution image generation with no $$$. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://ruoyidu.github.io/demofusion/demofusion.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English,

Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.03206.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://nips.cc/virtual/2023/poster/73566.

Xiefan Guo, Xinzhu Ma, Haiyu Zhang, and Di Huang. CTCal: Rethinking text-to-image diffusion models via cross-timestep self-calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43558–43567, 2026. URL https://arxiv.org/abs/2603.20741.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract.html.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2212.04488.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2301.07093.

Zhichao Liao, Xiaole Xian, Qingyu Li, Wenyu Qin, Meng Wang, Weicheng Xie, Siyang Song, Pingfa Feng, Long Zeng, and Liang Pan. PureCC: Pure learning for text-to-image concept customization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.07561.

Zhihang Lin, Mingbao Lin, Wengyi Zhan, and Rongrong Ji. AccDiffusion v2: Toward more accurate higher-resolution diffusion extrapolation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8351–8363, 2025. doi: 10.1109/TPAMI.2025.3576740. URL https://ieeexplore.ieee.org/document/11024232/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Brian B. Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Dynamic attention-guided diffusion for image super-resolution. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. doi: 10.1109/WACV61041.2025.00054. URL https://arxiv.org/abs/2308.07977.

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. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.01952.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.10752.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2208.12242.

Luigi Sigillo, Shengfeng He, and Danilo Comminiello. Latent wavelet diffusion for ultra-high-resolution image synthesis. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.00433.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.1556.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Unsplash. The Unsplash dataset. Dataset documentation, n.d. URL https://github.com/unsplash/datasets. Accessed September 13, 2026.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2101.04061.

Yu Xue, Haoxuan Qu, Zhuoling Li, Hongbin Xu, Jianxiong Yin, Simon See, Hossein Rahmani, and Jun Liu. HRDiT: Training-free high-resolution image generation with off-the-shelf diffusion transformer models. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2608.07003.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html.

A IMPLEMENTATION AND ACCOUNTING

Observation construction. DTD textures are converted to grayscale and resized to a 112×112 interior. For each aligned 16×16 observation block, the construction subtracts the mean over its intersection with that interior and uses a deterministic integer rounding adjustment that preserves an exact zero sum. Boundary intersections contain 64, 128, or 256 interior pixels. The residual is zero outside the interior and is scaled before being added to the shared carrier. Equality of L across details is checked before VAE encoding; Appendix N proves this observation invariance.

Fine-model optimization. The main SD1.5 configuration uses FP32 trainable parameters with BF16 autocast, rank-8 LoRA, and the coarse-conditioning input branch. AdamW uses $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 0.01. Four microbatches of size four give 16 scheduled slots per optimizer update. Losses are averaged over all scheduled slots, including those with zero supervision. The learning rate increases linearly to 10^{-4} over the first 100 updates. Caption dropout substitutes the frozen empty-caption embedding while leaving the coarse condition unchanged. Uniform and ROI train the same number of parameters.

Balanced support. The 32 retained photo–detail pairs are partitioned into eight four-element groups. The content group labeled "person" contributes two such groups; each of the six other content groups contributes one. With an ordering fixed before training, the layout assignment is a cyclic shift of the detail code by group rank. This yields eight supervised conditions per layout and two examples per detail/layout cell. Each person/layout cell contains two examples and every other content-group/layout cell contains one. Twenty-four of the 32 pairs change layout relative to the initial spread.

Automatic HF control. The HF mask is computed from a one-level Haar transform of the noisy latent. Squared LH/HL/HH energy is averaged across channels, interpolated to the latent grid, and normalized per image by min–max scaling. Denote this normalized energy map by $a(u)$. At diffusion timestep t, the binary mask selects locations satisfying $1000(a(u) + 0.3) \geq t$. The shifted control rolls the mask by half its height and width, preserving its histogram and sparsity while changing spatial alignment. It may still include useful detail. The fixed VAE and short continuation are substantive departures from a full LWD reproduction. All continuation arms start from the same parent checkpoint and use matched subsequent random streams.

Native-4K adaptation budgets. The two SD1.5 seed pairs are continuations from 512^2 checkpoints shared within each pair. Each arm adds 1,024 native-4096^2 optimizer updates and 2,048 native high-resolution visits, moving from absolute step 8,192 to 9,216. The legacy 16-slot schedule retains 14,336 zero-loss slots. The adapted parameters are rank-8 attention LoRA and the coarse-input convolution, totaling 1,605,888 parameters. These native visits are additional to the preceding 512^2 training and should not be read as the models' total lifetime exposure.

The concentrated SDXL pair uses a 2,048-update uniform common stage followed by 1,024-update uniform and ROI branches: 3,072 native updates and 6,144 high-resolution visits along either trajectory. The spread pair has the same update and visit counts. It starts from the same initial model, optimizer, and random-generator states as the concentrated run, but trains its own support-specific common stage; the two supports do not share a trained common parent. The spread support contains eight conditions per layout and two per layout/detail cell. Each update processes two supervised examples, with their summed loss divided by 16; this is not equivalent to the SD1.5 zero-slot random-stream schedule. The backbone comparisons therefore differ in training history, as well as in model and response instrument. We report within-pair changes rather than attributing cross-backbone differences to architecture alone.

Seed accounting. Each distributed and balanced 512^2 support design has two fine-model training replicates with different initialization and random-stream seeds. Both use the same evaluation coarse bank. The native-4K SDXL seed-1 fits use a fresh fine-model initialization and fixed seeds for their random streams; a 2,048-update common stage is followed by two 6,144-update branches, so the common stage plus either branch gives 16,384 high-resolution visits. They are the completed four-model set in the upper block of Table 4. The second seed of the eight-photograph support in Section 6.1, for both the position cell and the channel cells, was added after the first seed's results

under new protocols fixed before training; the first-seed protocols had ruled out further seeds. Its position-0 uniform and ROI fits serve as the reference for both, and both are reported in full whichever way they came out.

B ADDITIONAL PAIRED MEASUREMENTS

Table 3: **Texture-interior weighting partly repairs concentrated failure.** (w_T, w_F) give relative interior/frame coefficients with compensating background normalization. C, L, J are class, detected-location, and joint rates (%). Effects (pp) include 95% paired photograph-bootstrap CIs; all fits use seed 0. Outer is the original $\alpha = .5$ ROI. VAE mixing limits attribution to the frame alone.

Support	(w_T, w_F)	C	L	J	ΔJ vs. uniform	ΔJ vs. outer
Concentrated	$(8, 1)$	51.37	69.53	39.50	$+7.86$ $[4.25, 11.48]$	$+15.77$ $[12.50, 19.04]$
	$(1, 8)$	34.28	82.28	29.64	-2.00 $[-4.49, 0.63]$	$+5.91$ $[2.83, 9.13]$
Balanced	$(8, 1)$	37.16	91.50	34.38	$+7.28$ $[4.59, 9.96]$	-5.52 $[-8.59, -2.39]$
	$(1, 8)$	30.57	93.07	28.56	$+1.46$ $[-0.98, 3.76]$	-11.33 $[-14.01, -8.50]$

Table 4: **Native 4K: the support interaction does not replicate, and the unrequested-region side effect does.** The two blocks are *different models and are not comparable*. The upper block is the completed four-model set quoted in Section 6.3: SDXL, training seed 1, 8,192 updates from initialization, 16,384 high-resolution visits. The lower block is an earlier generation of 1,024-update continuations of a 2,048-step parent with 2,048 visits, whose smaller counts follow from that budget; its second count column is template detail presence. Counts are uniform → ROI over 16 photographs; each row's two intervals are the joint effect then the unrequested-region effect, from the same 2,000 paired photograph resamples. Interaction is balanced minus concentrated. All models score $0/64$ on ALL4, and response detectors differ by backbone, counting template responses rather than cyan frames.

Support	Joint (unif. → ROI)	Unreq. region	Δ joint, Δ region (pp)
SDXL four-model set, seed 1, 8,192 updates			
matched positions, 192 requests			
Concentrated	14 → 33	10 → 36	$+9.90$ $[+5.73, +13.55]$
			$+13.54$ $[+6.77, +20.31]$
Balanced	14 → 26	7 → 24	$+6.25$ $[+1.56, +10.94]$
			$+8.85$ $[+3.65, +14.06]$
Interaction			-3.65 $[-8.85, +1.56]$
			-4.69 $[-11.46, +2.08]$
all 256 requests			
Concentrated	25 → 47	13 → 45	$+8.59$ $[+5.47, +11.72]$
			$+12.50$ $[+6.64, +18.75]$
Balanced	21 → 36	7 → 29	$+5.86$ $[+1.95, +10.16]$
			$+8.59$ $[+3.91, +13.28]$
Interaction			-2.73 $[-7.42, +2.34]$
			-3.91 $[-8.98, +1.17]$
Earlier generation, 1,024-update continuations			
SD1.5 (seed 0), Concentrated	0 → 5	0 → 17	$+1.95$ $[+0.39, +3.91]$
			$+6.64$ $[+3.12, +10.56]$
SD1.5 (seed 1), Concentrated	0 → 5	0 → 17	$+1.95$ $[+0.00, +4.69]$
			$+6.64$ $[+1.56, +14.06]$
SDXL (seed 0), Concentrated	3 → 15	3 → 16	$+4.69$ $[+1.95, +7.42]$
			$+5.08$ $[+2.34, +7.81]$
SDXL (seed 0), Balanced	9 → 16	25 → 29	$+2.73$ $[-1.56, +7.03]$
			$+1.56$ $[-7.42, +10.94]$

B8 provides four details for each of eight background photographs, whereas B32 provides one detail for each of 32 backgrounds; both contain 32 high-resolution conditions. The B8 comparison changes support organization beyond position coverage, so the main claim relies on the B32 spatial

Table 5: **Geometric ROI gains are not recovered by the tested automatic masks.** All four continuations use the same parent and add 2,048 high-resolution visits. Per-request rates use 2,048 requests. Joint changes and paired photograph-bootstrap intervals are relative to uniform. HF denotes a high-frequency mask.

	Uniform	Automatic HF	Shifted HF	Geometric ROI
Class response (%)	42.24	40.77	41.65	49.41
Detected location (%)	92.53	92.19	92.14	92.29
Joint success (%)	**39.26**	**37.99**	**38.82**	**46.24**
ALL4 (/512)	8	5	9	13
Δ joint (pp)	—	-1.27	-0.44	$+6.98$
95% paired CI	—	$[-3.12, 0.63]$	$[-2.05, 1.32]$	$[4.15, 9.77]$

Table 6: **Paired effects across support and allocation controls.** Each effect is in percentage points, with its 95% photograph-bootstrap interval beneath. Location is the detected-location readout. ALL4 entries give reference $\rightarrow$ candidate counts out of 512. Numbers in parentheses identify fine-training seeds; single-position and B8 fits use seed 0. No seeds are pooled.

Comparison	Δ class	Δ location	Δ joint	ALL4
Support designs: ROI minus uniform				
B8: four details/photo	$+49.46$ $[46.14, 52.78]$	$+0.34$ $[-3.61, 4.39]$	$+44.53$ $[40.92, 48.10]$	$0 \rightarrow 175$
Single position 0	$+0.93$ $[-1.76, 3.81]$	-31.69 $[-36.72, -26.22]$	-7.91 $[-10.99, -4.44]$	$2 \rightarrow 8$
Single position 1	$+4.44$ $[1.03, 8.06]$	-25.98 $[-30.66, -21.29]$	-5.76 $[-9.72, -1.71]$	$5 \rightarrow 9$
Initial spread (0)	$+7.08$ $[3.81, 10.40]$	-3.17 $[-5.37, -1.07]$	$+5.32$ $[2.05, 8.64]$	$4 \rightarrow 5$
Initial spread (1)	$+21.63$ $[18.55, 24.56]$	-3.61 $[-5.76, -1.56]$	$+19.04$ $[15.87, 22.27]$	$0 \rightarrow 25$
Balanced spread (0)	$+13.96$ $[11.28, 16.65]$	-0.68 $[-3.96, 2.39]$	$+12.79$ $[9.72, 15.72]$	$1 \rightarrow 8$
Balanced spread (1)	$+18.21$ $[15.14, 21.19]$	-1.56 $[-3.61, 0.49]$	$+15.82$ $[12.94, 18.51]$	$1 \rightarrow 15$
Matched continuation: candidate minus reference				
Automatic HF $-$ uniform	-1.46 $[-3.32, 0.39]$	-0.34 $[-1.03, 0.29]$	-1.27 $[-3.12, 0.63]$	$8 \rightarrow 5$
Shifted HF $-$ uniform	-0.59 $[-2.15, 1.12]$	-0.39 $[-1.17, 0.39]$	-0.44 $[-2.05, 1.32]$	$8 \rightarrow 9$
ROI $-$ uniform	$+7.18$ $[4.44, 9.96]$	-0.24 $[-1.22, 0.73]$	$+6.98$ $[4.15, 9.77]$	$8 \rightarrow 13$
ROI $-$ automatic HF	$+8.64$ $[5.62, 11.52]$	$+0.10$ $[-1.12, 1.27]$	$+8.25$ $[5.37, 11.04]$	$5 \rightarrow 13$
ROI $-$ shifted HF	$+7.76$ $[5.08, 10.30]$	$+0.15$ $[-0.88, 1.17]$	$+7.42$ $[4.83, 9.81]$	$9 \rightarrow 13$

interventions. All rows in Table 6 use 2,048 requests and 512 quartets. The archived summaries also retain forced-location and invalid-output scores. Forced location uses the highest-scoring frame candidate without a presence threshold; detected location also requires that candidate to pass the threshold. Under balanced support, detected-location changes have intervals crossing zero, whereas forced-location changes are -3.13 points ($[-5.91, -0.59]$) and -2.78 points ($[-4.59, -1.03]$) for seeds 0 and 1. The small detected-location effects therefore do not imply that every spatial readout is unaffected. In the matched ROI continuation, the ALL4 effect is $+0.98$ points with interval $[-0.39, 2.34]$; the positive per-request effect does not establish a quartet gain.

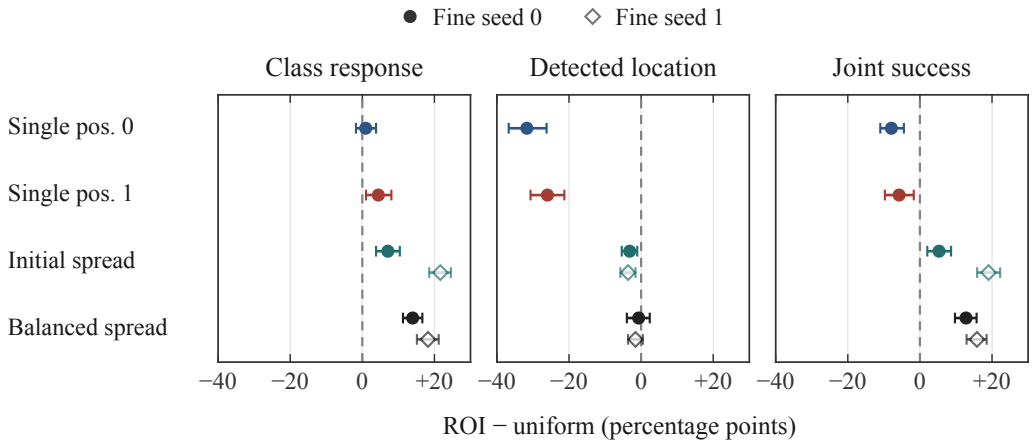

Figure 5: **Class, location, and joint effects under the reference configuration.** Points show ROI minus uniform at $\alpha = 1/2$, $\rho = 1/16$; bars are photograph-bootstrap 95% intervals. The columns share one scale. Fine seeds are offset within support groups. Joint success is the conjunction of class and location, so its effect is not their sum.

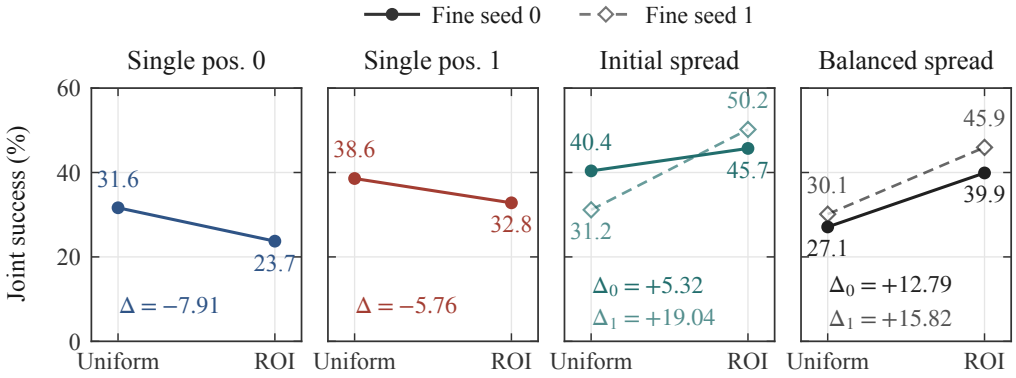

Figure 6: **The same weighting helps or hurts depending on where the detailed examples are.** Each line joins a uniform/ROI pair at $\alpha = 1/2$ and $\rho = 1/16$; panels share a joint-success scale and marker styles distinguish fine seeds. Labels give percentages and paired changes in points. Table 2 reports the intervals.

Automatic masks against the geometric ROI. Every result above marks the region with a visible frame. As an alternative we compare 1,024-update continuations from a common parent using an online Haar high-frequency (HF) mask of the noisy latent, its shifted copy, or uniform loss (Table 5). Only geometric ROI has a positive joint effect whose interval excludes zero, exceeding automatic HF by $+8.25$ points ($[5.37, 11.04]$) and shifted HF by $+7.42$ ($[4.83, 9.81]$). The shift preserves mask histogram and sparsity while changing alignment, but comparisons against uniform change sparsity and loss mass too, so alignment is not isolated.

C Readout history and interpretation

The selected classifier uses four one-vs-rest SVMs with penalty parameter $C = 1$ and frozen VGG descriptors. Its confusion matrix on the confirmation set, with true classes in rows and predicted

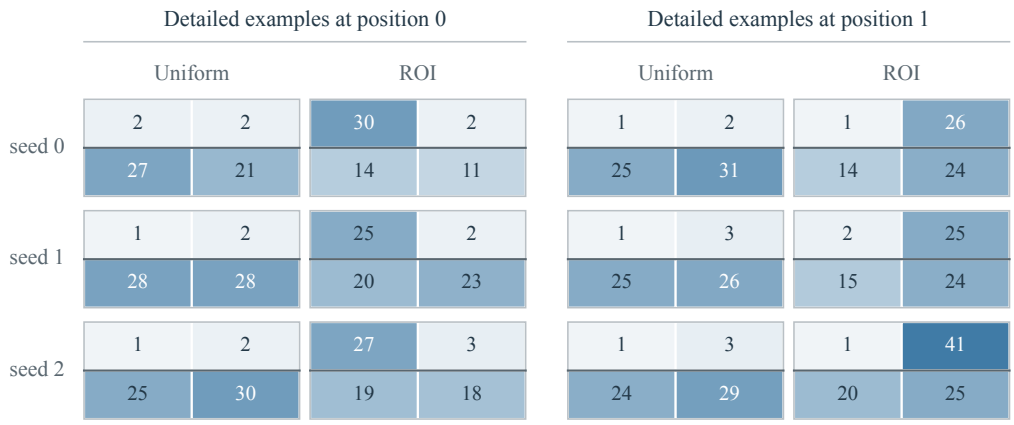

Figure 7: **Detections laid out on the geometry the generator writes into.** Each map is one model; its four cells are the four panel positions in their actual arrangement, shaded and labelled by how many of the 64 requests carry a qualifying frame there. Requests sit in the two lower positions, so the lower cells are the requested panel and the upper cells are never requested. Under ROI weighting the occupied upper cell moves with the training position in all three seeds, while the lower cells and the whole uniform column stay where they are. Seeds are not pooled.

Table 7: Readout development and confirmation. These are instrument measurements, not generator performance.

Instrument	Evidence split	Correct / total	Acc. (%)	Decision
Earlier readout	Confirmation	23 / 32	71.88	Gate failed
Earlier readout	Fresh confirmation	42 / 64	65.63	Gate failed
Replacement A	Validation	126 / 156	80.77	Selected
Replacement B	Validation	125 / 156	80.13	Not selected
Frozen A	Confirmation	45 / 60	75.00	Gate met exactly

classes in columns, is

	woven	knitted	meshed	marbled
woven	9	3	2	1
knitted	3	12	0	0
meshed	0	0	12	3
marbled	0	2	1	12

. (7)

All confirmation inputs were valid. Errors vary by class, and the classifier has no reject class for absent or implausible texture. Its 75% confirmation accuracy cannot provide a single scalar correction for generator scores, since classifier errors may depend on image content and generator artifacts. The remaining reserved test examples were not used to select or refit the classifier.

Prospective versus retrospective evidence. Results obtained with the replacement classifier are reported as a reanalysis. The original metric definitions and preregistered pass/fail outcomes are retained separately and are not relabeled using the new classifier. Under the replacement readout, the initial-spread seed-0 ALL4 effect has an interval that includes zero; it does not establish improved four-way control for that fitted pair.

D QUALITATIVE EVIDENCE AND SELECTION

The visualizations use recorded model outputs, with PNG hashes checked against the sampling records. Model comparisons share the coarse condition and the fine generator's sampling noise. Full

views retain the entire image extent; detail crops use identical coordinates across models. Generated pixels are not sharpened, recolored, or contrast-enhanced. The coarse input in Figure 13 is a display conversion of the stored float32 observation, not the raw coarse-generator decode. Class references come from the 32 permitted single-position training conditions, using the lowest allowed array index in each class.

The training-position panel. Figure 1 uses request `ll_test_b010_l2_d2_n0`, requested at position 2. Among the 64 main-analysis requests we keep those where the frozen geometric rule finds the requested frame in all four models, exactly one extra frame at the trained position under ROI in both training arms, and no frame at either upper position under uniform. Six requests qualify and we show the lexicographically first; no score was consulted. The carets lie outside the image area and touch no generated pixel. Per-model source hashes are in `evidence/qualitative_position_provenance.json`.

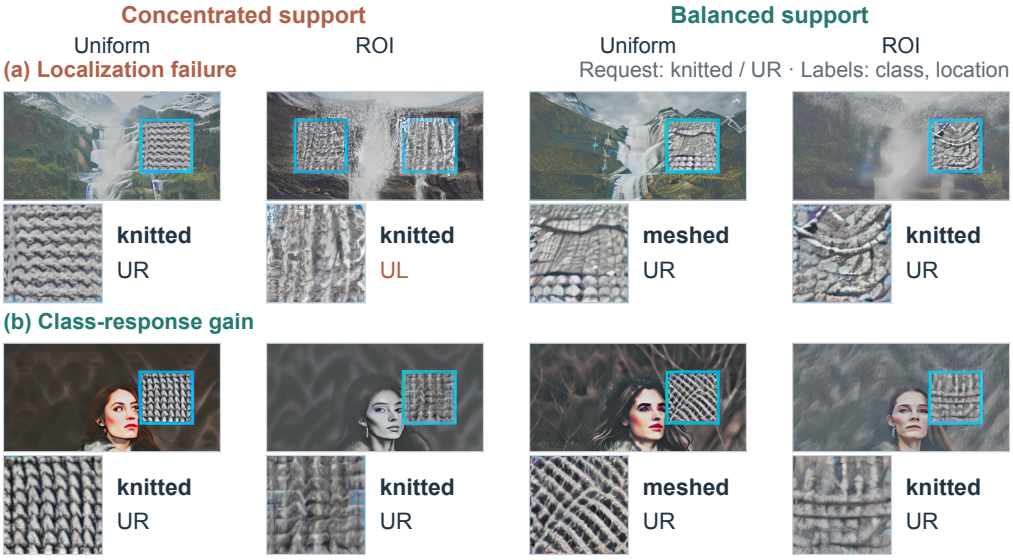

Figure 8: **Local weighting changes both where and what a model generates.** Two knitted/upper-right requests, matched on input and noise across four models. Each view shows the source image's upper half and a 112 × 112 crop of the requested detail region, preserving the original pixels. (a) Concentrated ROI retains a knitted response but adds an upper-left frame, at the position its high-resolution examples occupied. (b) Balanced ROI changes the class prediction from meshed to the requested knitted class. These are post-hoc illustrations (fine seed 0).



Figure 9: **Every request that meets the disclosed criterion, not a chosen example.** All six qualifying requests under the ROI model trained at position 0 (top) and at position 1 (bottom); the requested position is given above each column. In each case the requested panel stays where it was asked for and the second panel moves between the two unrequested upper positions. Original pixels, no crop or adjustment; the criterion and hashes are in Appendix D.

Illustrative 512 requests. Figure 8a uses request `b004/l1/d1/n1`, requesting knitted texture at upper right. Among the 512 upper-right requests, 23 have joint success for single-position uniform

and a correct class but incorrect location for single-position ROI, whose predicted location is upper left. We select the lexicographically first of these 23 and show all four model arms. Balanced-model outcomes were not used to select this case. Figure 8b uses `b000/l1/d1/n0`, the lexicographically first of 149 upper-right requests on which balanced uniform has correct detected location but an incorrect class response, while balanced ROI succeeds jointly. All four model arms are again retained. Both cases were selected post hoc to illustrate behavior; neither estimates its prevalence.

Complete 512 quartet. Figure 14 contains all four detail requests at the lowest background and noise indices (background 0, layout 1, noise 0), without filtering on their outputs. This index-based selection was made for presentation after the experiments; it was not a prospective visual evaluation set. The requested class appears at left, while the labels below generated images are classifier and location predictions. Repeated knitted predictions across requests expose the remaining control failures.

Crop and reference definitions. Main Figure 8 uses the exact upper-half box $[0, 0, 512, 256]$ for spatial context; Figure 13 provides the complete images. For 512 outputs, the requested upper-right detail uses the half-open pixel box $[328, 72, 440, 184]$, and the unrequested upper-left region uses $[64, 64, 192, 192]$. Training references use their own upper-left detail box $[72, 72, 184, 184]$. They show supervised details from different source photographs and are not paired ground truth for the generated evaluation images.

Native 4K visual provenance. The SDXL visual set was fixed before generation: four photographs, each with all four detail requests at layout code 1. Figure 10 shows the complete quartet for photograph 0, layout code 1 (physical upper left), and noise repetition 0. This choice is based on the photograph index; no detail request is omitted. The full-view panel shows d0 and d3, selected retrospectively to contrast an undetected request and the rare detected response. Figure 15 shows complete images for all four requests. Uniform has no detected response in this quartet; ROI has one, for d3. Across the full evaluation, uniform and ROI have detected responses in 3/256 and 16/256 requests, respectively.

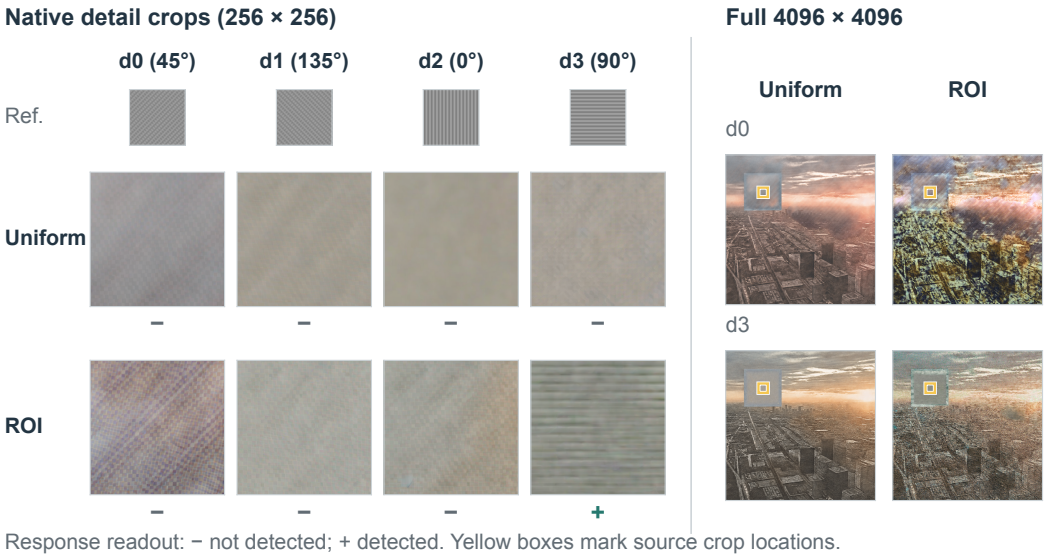

Figure 10: **Four native 4K requests, only one detected ROI response.** In the concentrated SDXL pair, each column retains the same 256^2 source-pixel crop for uniform and ROI. Development references identify the patterns; angles denote wave-vector direction. The d0/d3 full views expose global artifacts; all scenes appear in Appendix G. The quartet is fixed by index; the two overview requests were chosen retrospectively. Detector labels describe response, not fidelity.

Every 4K detail crop uses the half-open source-pixel box $[896, 896, 1152, 1152]$ inside the requested panel $[512, 512, 1536, 1536]$. Full views are downsampled only for display and retain the surrounding

scene, including noise and color artifacts. The reference columns are crops of existing development fixtures, not paired evaluation targets. Angles denote the directional pattern's wave vector, not the visible stripe orientation. Crop pixels are unchanged; no contrast adjustment, sharpening, or learned enhancement is applied.

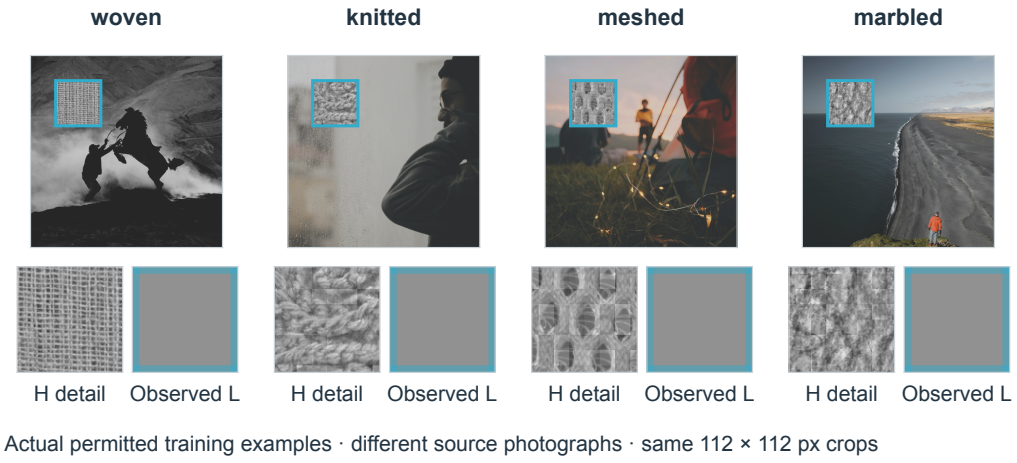

Figure 11: **Actual observed training details and their coarse counterparts.** Each column is one permitted high-resolution condition from a different training photograph. The lower panels display the same 112×112 region in the target and its low-resolution observation. The grayscale texture distinction is absent from the latter; the frame supplies a location cue. These are constructed training targets, not generator outputs.

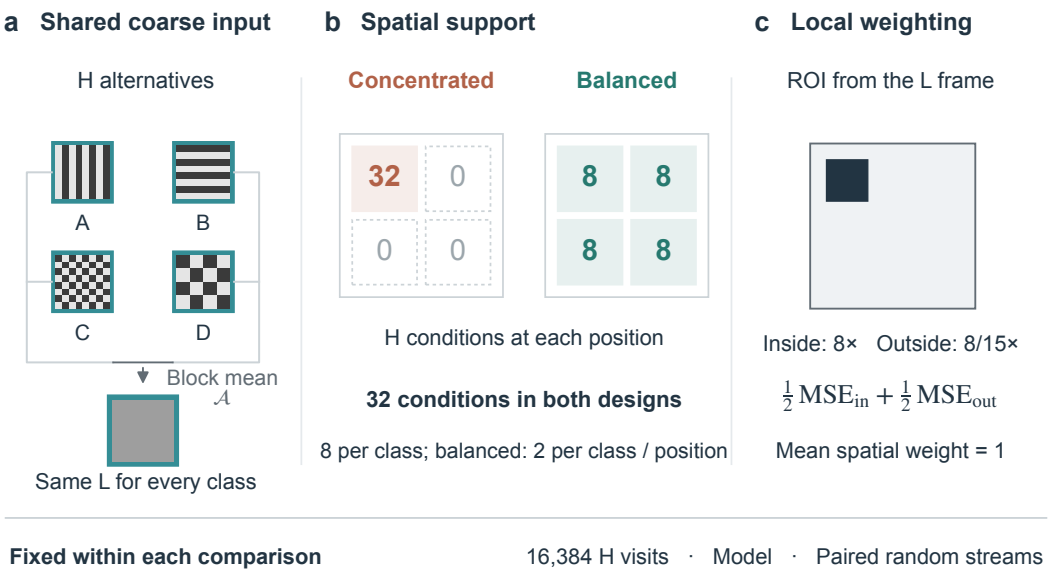

Figure 12: **The same detail budget can have different spatial support.** (a) Four detail targets share one low-resolution observation: the frame remains visible, while the texture distinction is removed. (b) Concentrated and balanced supports both contain 32 photo–detail pairs. Balanced support assigns eight pairs to each position, including two from each detail class. (c) ROI weighting assigns half the loss mass to the framed region, whose latent area is $1/16$ of the grid. Patterns A–D are schematic alternatives, not generated outputs or DTD exemplars.

E FULL CONTEXT FOR THE PAIRED 512 EXAMPLES

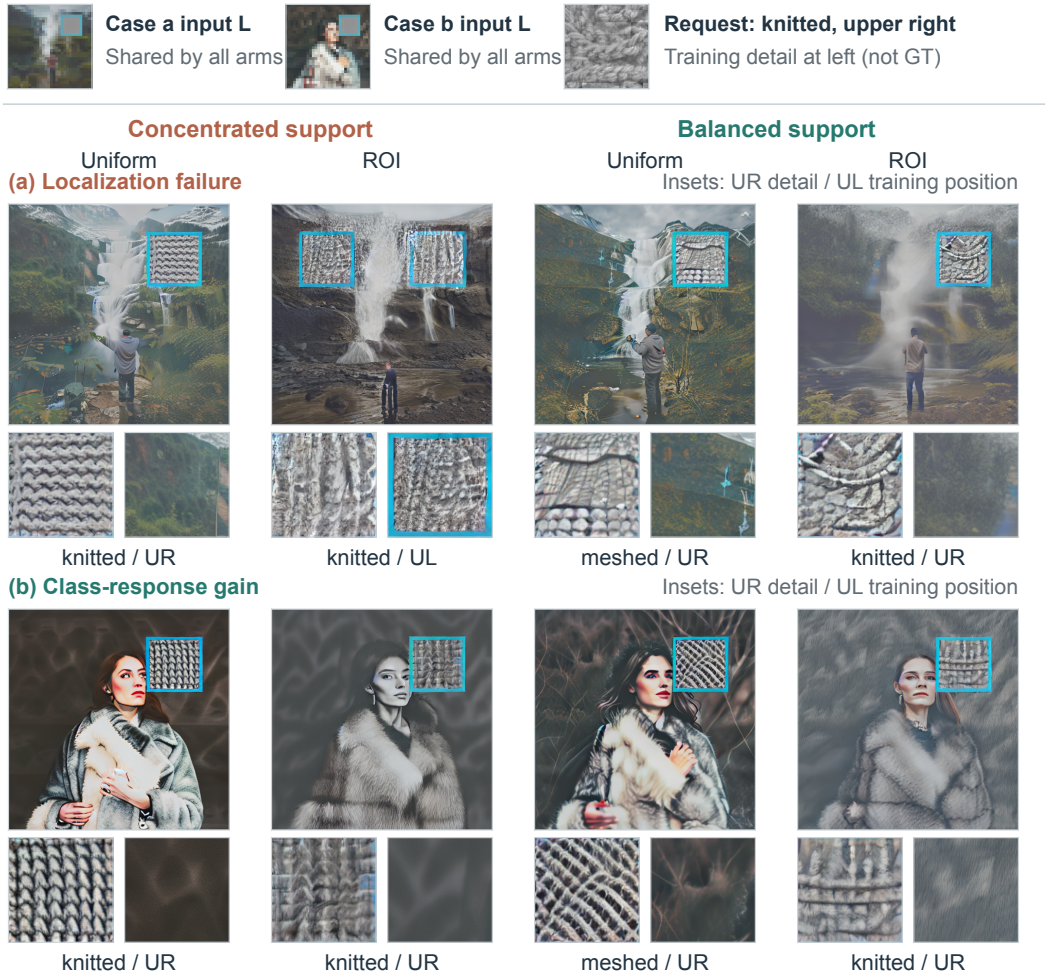

Figure 13: **Complete scenes behind the compact main-text comparisons.** The two requests and all four model arms match Figure 8. Full 512^2 images retain the surrounding scene; paired insets show the requested upper-right detail and the unrequested upper-left region. The shared L inputs and a permitted training reference from a different photograph are shown above. Class/location labels are readout predictions, not fidelity judgments. Selection is described in Appendix D.

F COMPLETE 512 GENERATION QUARTET

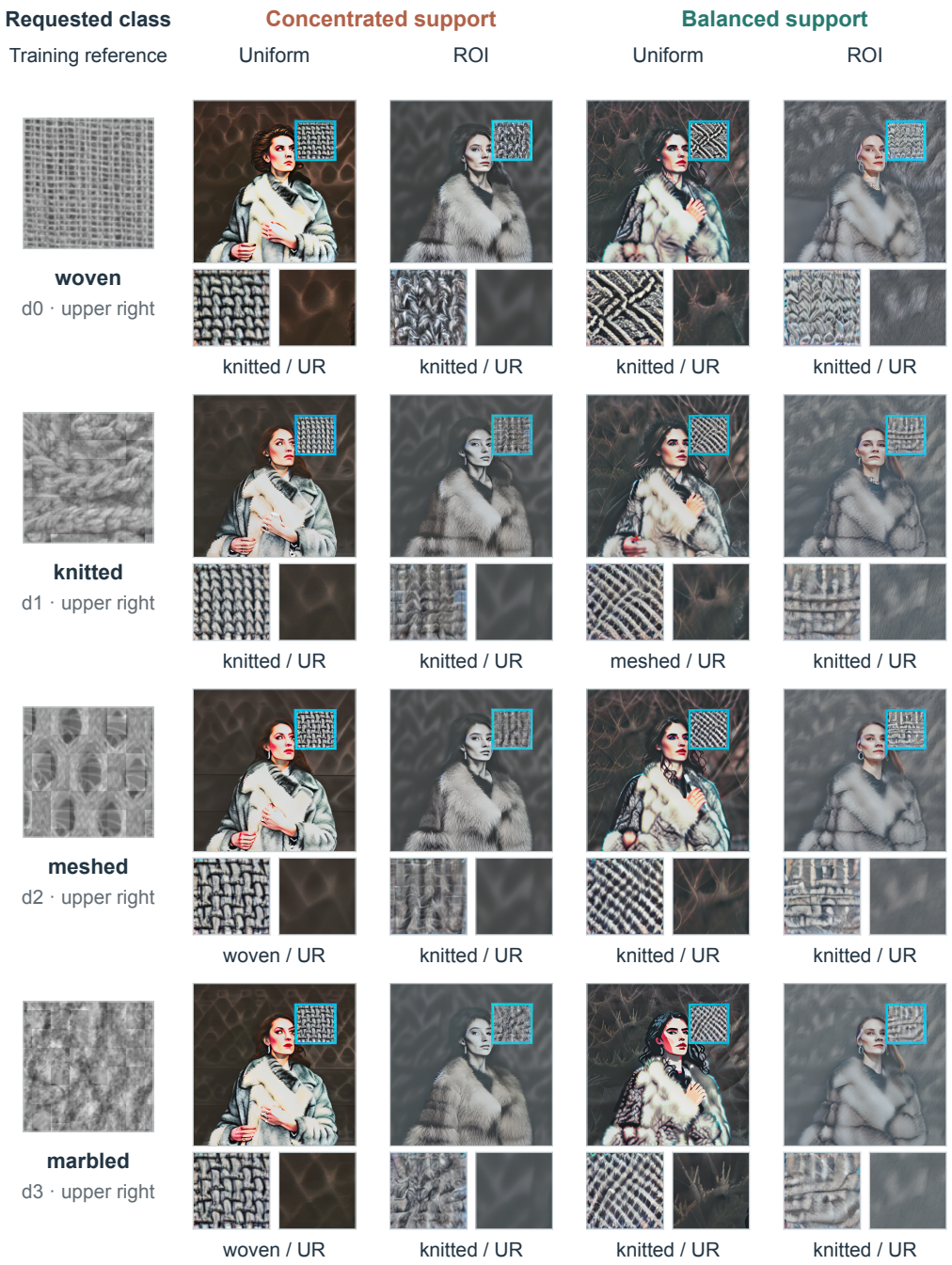

Figure 14: **All four requests are retained.** Background 0, upper-right layout, and noise repetition 0 under each of the four fine-seed-0 models. Each output is accompanied by its upper-right detail crop and upper-left region crop. Labels below images give class/location predictions; the left column shows permitted training references from different photographs for the requested class. This single quartet illustrates control failures and is not a success-rate estimate.

G FULL NATIVE 4K QUARTET

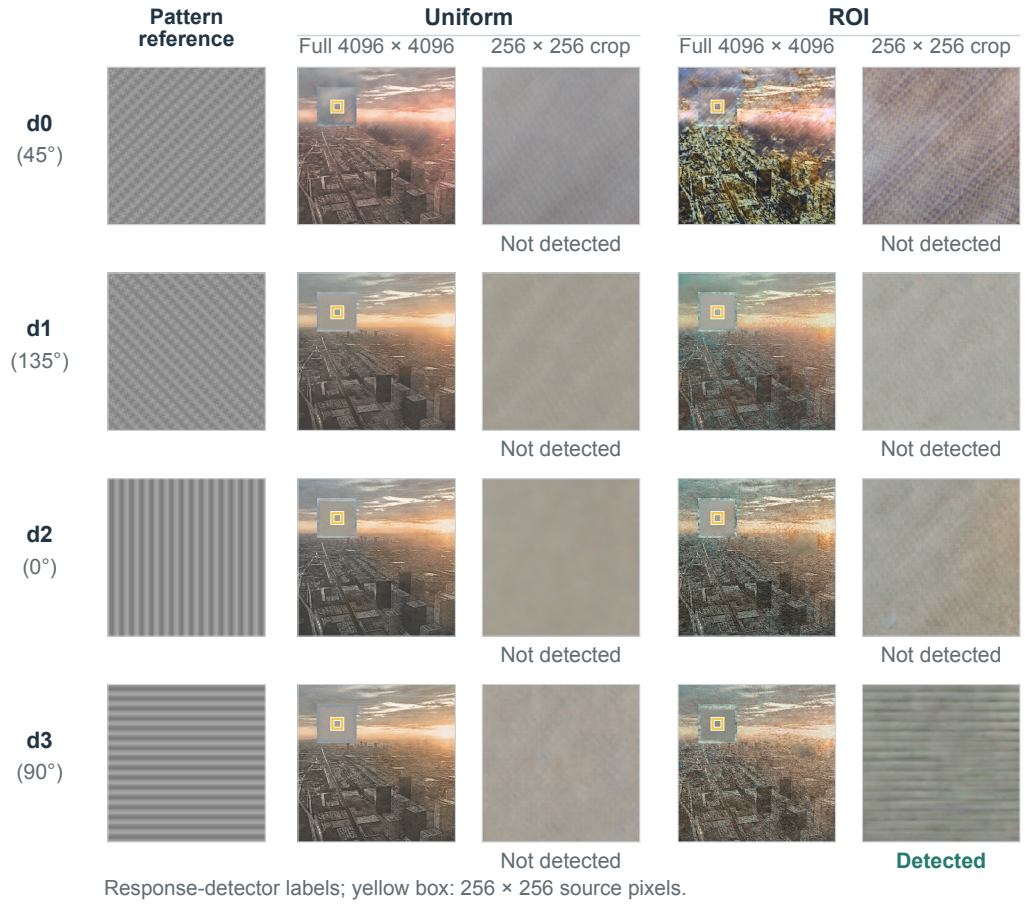

Figure 15: **Every full scene is retained for the fixed native-4K quartet.** Four requests at photograph 0, layout code 1 (physical upper left), and noise repetition 0, under uniform and ROI. Yellow boxes mark the common unchanged 256^2 detail crop. Uniform has no detected response in this quartet; ROI has one, for d3. Full views retain noise and color artifacts. References are development fixtures, not paired ground truth.

H THE POSITION-INHERITANCE AUDIT, IN FULL

The contribution of this paper is a controlled measurement, so the measurement itself has to be reconstructible. This appendix states the request selection, the stratum composition, every threshold in the frozen geometric rule, the six frame-free descriptors, and the resampling, at the level needed to reimplement them. All of it was fixed before the two replication seeds were trained, and the values below are transcribed from the frozen protocol record rather than restated from memory.

How the audited requests are selected. Each 512^2 model is evaluated on 2,048 requests: 64 held-out backgrounds $\times$ 4 requested layouts $\times$ 4 requested detail classes $\times$ 2 noise repetitions. The audit uses a subset chosen by a deterministic function of the background index alone, with no reference to any image, score, or detection. For background b, write

$$\begin{aligned} q_0(b) &= \big(b \bmod 4,\ \lfloor b/4 \rfloor \bmod 4,\ \lfloor b/16 \rfloor \bmod 2\big), \\ q_1(b) &= \big(q_0^{(1)}+2 \bmod 4,\ \ q_0^{(2)}+1 \bmod 4,\ \ 1-q_0^{(3)}\big), \end{aligned} \tag{8}$$

read as (requested layout, requested detail class, noise repetition). A request is audited when its triple equals $q_0(b)$ or $q_1(b)$. This yields exactly 128 requests per model: two per photograph, four in each of the 32 (layout, detail, noise) cells. The selection is recorded as containing no score-dependent step, and the same 128 request identities are used for every model in every seed.

The primary stratum is the 64 of those requests whose *requested* layout is 2 or 3, so that panel positions 0 and 1 are unrequested in every audited image and Q is measured only where nothing was asked for. Its composition is fixed by Equation 8 and is not balanced by hand: 32 requests at layout 2 and 32 at layout 3, 16 in each of the four detail classes, 32 in each noise repetition, and exactly one request from each of the 64 photographs. The remaining 64 audited requests, whose requested layout is 0 or 1, are reported as the all-128 stratum together with the primary one; they are not a separate experiment. Every audited image is bound to its model by eleven recorded fields, including the sample identity, the PNG hash, the noise hash, the coarse latent row hash, the observed-RGB row hash and the public-rows hash, all of which must match between the sampling record and the evaluation record before the image is scored.

The frozen geometric rule. A candidate cyan region is the set of pixels satisfying, in 8-bit RGB,

$$\min(G,B) \geq 100, \qquad \min(G,B) - R \geq 35, \qquad |G-B| \leq 100, \tag{9}$$

after a 5×5 binary closing and 8-connected labelling. A component is considered only if its bounding box has width and height in $[80, 180]$ pixels, an aspect ratio in $[0.6, 1.667]$, and at least 64 cyan pixels. Inside that box, with a border band of 12 pixels, let s_k be the cyan fraction of band k, let c_k be the fraction of that band's long axis covered by at least one cyan pixel, let ι be the cyan fraction strictly inside the bands, and let ϕ be the share of the component's cyan pixels lying in the bands. The component *qualifies* as a frame when

$$\#\{k : s_k \geq 0.10 \ \wedge \ c_k \geq 0.75\} \geq 3, \qquad \phi \geq 0.5, \qquad \iota < 0.20. \tag{10}$$

A component that passes the size filter but fails Equation 10 is recorded as an uncertain candidate rather than silently dropped, and those counts are reported. A qualifying frame is assigned to the panel whose centre is nearest its own, and is called canonical only when its centre lies within a tolerance of 48 pixels of that panel centre in both coordinates; frame_j counts canonical frames at position j. Non-detection is not evidence of absence, and nothing in Equations 9–10 refers to texture.

The six frame-free descriptors. For each panel position, the 128×128 crop is converted to luminance $g \in [0,1]$ by $0.299R + 0.587G + 0.114B$. Two descriptors are computed directly: $\mathrm{gray_std}$ is the standard deviation of g, and $\mathrm{gradient_energy}$ is the mean of the squared first differences along the two axes, averaged. The remaining four use the power spectrum of the mean-removed crop under a separable Hann window, $p = |\mathcal{F}[(g-\bar{g})\,w]|^2$ with $w_{ij} = \mathrm{hann}_i\,\mathrm{hann}_j$. With normalized radial frequency ρ, let $H = \{\rho \geq 0.15\}$ be the high band, $V = \{\rho \geq 1/n\}$, and $m = \sum_H p$. Then $\mathrm{hf_energy} = m/\Big(n^4\overline{w^2}\Big)$, $\mathrm{hf_fraction} = m/\sum_V p$, the directional anisotropy is $\big|\sum_H p\, e^{2i\theta}\big|/m$ with θ the spectral angle, and the spectral entropy is the Shannon entropy of p/m over H divided by $\log|H|$. All six are reported for every contrast; none was selected after a result

was seen. They summarize how much structure sits at a position, not which texture class is there, and the frame still contributes to them, so they remove the readout's dependence on frame detection and not the frame itself.

Estimands and resampling. With $Q_a = \mathrm{frame}_1 - \mathrm{frame}_0$ summed over the stratum and divided by its size, $M(w) = Q_{h=1,w} - Q_{h=0,w}$ and $K = M(\text{ROI}) - M(\text{uniform})$. Because Q spans $[-100, 100]$ points, K spans $[-400, 400]$. Intervals are percentile bootstrap over photographs, resampling whole photograph blocks so that all four fits of a seed move together: 2,000 replicates from `PCG64(731031)`, drawn once as a 2000×64 integer array whose SHA-256 is recorded in the protocol and re-verified by every readout before it computes anything. Seeds are reported separately and never pooled, and the interval describes variation across held-out photographs for fixed fitted models and a fixed rule; it does not include training-seed variation.

I TARGETED WEIGHTING INTERVENTIONS AND OUTPUT DIAGNOSTICS

I.1 TWELVE ADDITIONAL FITS UNDER A COMMON PROTOCOL

The follow-up matrix crosses two support designs—single position 0 and balanced spread—with six new weighting configurations. All twelve fits use fine-training seed 0 and start from the original initialization. They retain the same 32 photo–detail source pairs, 8,192 optimizer updates, 16,384 high-resolution visits, and 114,688 scheduled zero-loss slots. Within each support, initialization, training examples, random streams, and optimization settings are shared with the reference fits. Across supports, the layout reassignment changes targets and conditions as described in Section 6.1. The existing uniform and reference ROI models are reused.

Evaluation uses the same 2,048 requests, shared coarse samples, and paired sampling noise as the original comparisons. The class classifier and frame detector remain frozen. The full matrix reports each new model against both uniform and the reference ROI configuration; intervals use the photograph bootstrap in Section 4. These are targeted comparisons for one training seed, with marginal intervals for individual contrasts rather than a simultaneous confidence band. No usable human evaluation of generated outputs was obtained (Appendix M).

Texture-interior and frame-ring coefficients. On the 64×64 latent grid, let T be the geometric 14×14 texture interior, F the surrounding ring within the 16×16 panel, and O the remaining grid. Their sizes are 196, 60, and 3,840. For relative weights (w_T, w_F, w_O), define the supervised loss

$$\mathcal{J}(w_T, w_F, w_O) = \mathbb{E}\left[\frac{w_T \sum_{u\in T} e_\theta(u) + w_F \sum_{u\in F} e_\theta(u) + w_O \sum_{u\in O} e_\theta(u)}{4{,}096}\right],$$
$$196w_T + 60w_F + 3{,}840w_O = 4{,}096. \tag{11}$$

The constraint preserves mean spatial weight one. Table 8 gives the four coefficient combinations, including the two existing references. The outside compensation is part of each intervention. Targets, visible frames, and model inputs remain unchanged; VAE receptive fields can also mix information across these geometric regions. Reducing the frame coefficient therefore does not remove frame information or identify a purely semantic texture effect.

Table 8: **Normalized region coefficients.** Weights are relative to the uniform spatial mean. Only the last two configurations require new fits for each support.

Configuration	Texture w_T	Frame w_F	Outside w_O
Uniform	1	1	1
Reference ROI ($\alpha = 1/2$)	8	8	$8/15$
Texture coefficient eight	8	1	$617/960$
Frame coefficient eight	1	8	$57/64$

Loss mass and footprint. For a fixed 16^2 mask, $\alpha \in \{1/4, 1/2, 3/4\}$ gives inside weights $4, 8, 12$ and outside weights $4/5, 8/15, 4/15$, respectively. The two nonreference values add four fits across supports. The remaining four fits use centered 12^2 and 20^2 masks at $\alpha = 1/2$, with the original targets held fixed. Their area fractions are $9/256$ and $25/256$, compared with $1/16$ for the reference mask.

Mask size. At fixed $\alpha = 1/2$, balanced effects stay positive for 12^2, 16^2 and 20^2 masks ($+12.74$, $+12.79$, $+12.01$), while concentrated effects are -9.77, -7.91 and $+4.54$. Enlarging the mask can thus recover a positive effect without adding supervised positions, and a second seed reproduces the weakened interaction at 20^2. These fits use one seed except where a second is stated. The gain shows up per request rather than per quartet: balanced reference fits solve 8/512 and 15/512 quartets under ROI, against 1/512 under uniform (Appendix B). Of the automatic masks tested, only the geometric ROI has a positive joint effect excluding zero (Appendix B). Inside/outside weights are $(128/9, 128/247)$ for 12^2, $(8, 8/15)$ for 16^2, and $(128/25, 128/231)$ for 20^2. The smaller mask covers part of the nominal texture interior, while the larger one includes the frame and

some surrounding area. Thus footprint, region coverage, and per-location coefficient change together. This experiment tests the joint change; it cannot attribute the effect to area alone.

Table 9: **All 12 completed loss-weighting interventions.** Rates C, L, J use 2,048 requests from 64 photographs per model. Effects are paired joint differences (pp), with 95% photograph-bootstrap intervals on the second line. All fits use training seed 0 and the original 8,192-update, 16,384-high-resolution-visit budget. The first two settings reweight the texture interior and frame ring; other settings change ROI loss mass α or its latent footprint. All weights have spatial mean one. Footprint changes hold the image target fixed and jointly change per-site weight and covered content. Outer denotes the original 16×16, $\alpha = .5$ ROI; reused references are not new fits.

Loss setting	C (%)	L (%)	J (%)	ΔJ vs. uniform	ΔJ vs. outer
Concentrated: uniform $J = 31.64\%$; outer ROI $J = 23.73\%$					
$w_T = 8,\ w_F = 1$	51.37	69.53	39.50	$+7.86$ $[4.25, 11.48]$	$+15.77$ $[12.50, 19.04]$
$w_T = 1,\ w_F = 8$	34.28	82.28	29.64	-2.00 $[-4.49, 0.63]$	$+5.91$ $[2.83, 9.13]$
$16\times16,\ \alpha = .25$	44.29	83.69	38.48	$+6.84$ $[3.56, 10.01]$	$+14.75$ $[11.96, 17.48]$
$16\times16,\ \alpha = .75$	41.02	37.11	19.43	-12.21 $[-15.09, -9.33]$	-4.30 $[-6.98, -1.56]$
$12\times12,\ \alpha = .5$	39.21	46.09	21.88	-9.77 $[-12.50, -7.08]$	-1.86 $[-4.59, 0.73]$
$20\times20,\ \alpha = .5$	45.46	74.46	36.18	$+4.54$ $[1.07, 8.11]$	$+12.45$ $[9.42, 15.63]$
Balanced: uniform $J = 27.10\%$; outer ROI $J = 39.89\%$					
$w_T = 8,\ w_F = 1$	37.16	91.50	34.38	$+7.28$ $[4.59, 9.96]$	-5.52 $[-8.59, -2.39]$
$w_T = 1,\ w_F = 8$	30.57	93.07	28.56	$+1.46$ $[-0.98, 3.76]$	-11.33 $[-14.01, -8.50]$
$16\times16,\ \alpha = .25$	36.43	91.60	33.98	$+6.88$ $[4.44, 9.23]$	-5.91 $[-8.98, -3.08]$
$16\times16,\ \alpha = .75$	36.87	84.23	32.28	$+5.18$ $[2.88, 7.76]$	-7.62 $[-10.16, -5.08]$
$12\times12,\ \alpha = .5$	44.09	90.04	39.84	$+12.74$ $[10.01, 15.58]$	-0.05 $[-2.98, 2.98]$
$20\times20,\ \alpha = .5$	41.80	92.33	39.11	$+12.01$ $[9.86, 14.36]$	-0.78 $[-3.52, 2.20]$

I.2 Separating two frame-localization failures

The original location readout computes four frame scores s_j, chooses $\ell^\star = \arg\max_j s_j$, and declares detected localization correct when $\ell^\star$ equals the requested position ℓ and $s_{\ell^\star}$ exceeds the frozen threshold τ. Forced localization omits this threshold check. Detected-localization failure therefore has the disjoint partition

$$\{s_\ell \le \tau\} \ \dot\cup\ \{s_\ell > \tau,\ \ell^\star \ne \ell\}. \tag{12}$$

The first event means that the requested candidate fails the detector threshold; it does not establish physical frame absence. The second means that another candidate wins despite the requested candidate passing. An extra detected candidate is insufficient to cause failure unless it wins. The texture classifier always uses the requested crop, independently of the winning frame position.

Both components increase under concentrated support. Under balanced support, fewer requested candidates fail the threshold but more lose to a competitor; the small aggregate localization change conceals these opposing movements. The association also follows the selected training position: with high-resolution support at position 1, that position wins on 32 versus 516 of the 1,536 requests for other positions. These counts describe the forced argmax, irrespective of its threshold status. This post hoc partition reuses stored scores and the original threshold without refitting. It characterizes frame competition, while leaving texture presence, displacement, and fidelity unresolved.

Table 10: **An exact partition of detected-localization failures.** Entries give uniform → reference ROI counts out of 2,048 requests per model. The two columns are mutually exclusive and sum to the original localization-failure count. Single-position fits use seed 0; parentheses identify balanced-support fine-training seeds.

Support	Requested frame fails threshold	Requested frame passes, competitor wins
Single position 0	$267 \to 637$	$77 \to 356$
Single position 1	$128 \to 348$	$77 \to 389$
Balanced spread (0)	$97 \to 51$	$97 \to 157$
Balanced spread (1)	$80 \to 46$	$89 \to 155$

I.3 Additional diagnostics for the native-4K SDXL spread pair

The completed SDXL spread comparison uses fine-training seed 0, 3,072 updates, and 6,144 high-resolution visits per model. All four layout codes occur in high-resolution training support. Its 256 evaluation requests use 16 held-out photographs and form 64 quartets. Table 11 separates forced direction, detected response, and extra-template flags. The paired intervals resample the 16 photographs, with 2,000 draws.

Table 11: **SDXL spread diagnostics at native** 4096^2**.** Entries are observed counts; changes and 95% paired intervals are in percentage points. Forced direction rises from 32.03% to 50.00%, while the response and joint intervals include zero. The zero-width empirical ALL4 interval reflects all-zero observations, not certainty about its population rate.

Readout	Uniform	ROI	Δ	95% interval
Forced direction	$82/256$	$128/256$	$+17.97$	$[10.94, 24.61]$
Detected response	$25/256$	$29/256$	$+1.56$	$[-7.42, 10.94]$
Joint success	$9/256$	$16/256$	$+2.73$	$[-1.56, 7.03]$
Any extra-template flag	$177/256$	$204/256$	$+10.55$	$[4.30, 18.36]$
ALL4	$0/64$	$0/64$	0.00	$[0.00, 0.00]$

Forced direction assigns a class even when the output fails the detected-response rule. Extra-template flags test the four fixed candidate geometries; they are not a validated count of arbitrary unwanted objects. Neither diagnostic certifies correct texture placement or whole-image fidelity. The spread pair supplies a completed within-support comparison on new photographs, but does not independently replicate a cross-resolution interaction between support and weighting.

J DIRECT INTERACTIONS AND SENSITIVITY TO THE LOCALIZATION RULE

These post-hoc analyses use the completed seed-0 models. They address two separate questions: whether the weighting effect differs between supports, and how that conclusion depends on the frame-competition rule.

J.1 PAIRED SUPPORT AND CONFIGURATION CONTRASTS

Let B denote balanced support, S support concentrated at position 0, U uniform weighting, and w_0 the reference 16^2 ROI with $\alpha = 1/2$. For measured joint success J, define

$$\begin{aligned} I(w) &= (J_{B,w} - J_{B,U}) - (J_{S,w} - J_{S,U}), \\ D(w) &= I(w) - I(w_0) = (J_{B,w} - J_{B,w_0}) - (J_{S,w} - J_{S,w_0}). \end{aligned} \tag{13}$$

Positive I means that weighting is more beneficial under balanced support. D directly tests how this support difference changes with the weighting configuration; the uniform terms cancel.

All four models in a contrast share 2,048 request identities and the same 64 photographs, with 32 requests per photograph. We verified the photograph, layout, detail, and repetition keys, the stored coarse-input and fine-noise hashes, and the frozen readout provenance. Each of 2,000 bootstrap draws resamples the same photographs jointly across all models; all settings reuse this draw matrix (seed 731031). A second implementation verifies the estimates and percentile endpoints from integer event counts. These intervals describe photograph variation for fixed fits and a fixed instrument, excluding training-seed and readout uncertainty.

Table 12: **The support interaction in all three training seeds.** Joint successes out of 2,048 requests for each of the four fits per seed, the ROI-minus-uniform difference within each support, and their difference. Counts are recounted from each model's own evaluation records and reproduce the reported interaction in every seed. Intervals are paired photograph bootstrap over 64 photographs; seeds are never pooled. The balanced gain replicates in all three seeds while the concentrated effect does not, which is why we report the interaction rather than a sign reversal.

Seed	Concentrated		Balanced		Δ (pp)		Interaction (pp), 95% CI
	Unif.	ROI	Unif.	ROI	Conc.	Bal.	
0	648	486	555	817	−7.91	+12.79	$+\mathbf{20.70}$ [+16.21, +25.00]
1	563	639	617	941	+3.71	+15.82	$+\mathbf{12.11}$ [+8.30, +15.92]
2	568	614	603	770	+2.25	+8.15	$+\mathbf{5.91}$ [+1.95, +9.67]

Table 13: **Direct support interactions and their changes across weighting settings.** $I(w)$ is the balanced-minus-concentrated difference in weighting effects on joint success; $D(w) = I(w) - I(w_0)$ compares this interaction with the reference ROI. Effects are percentage points, with 95% paired photograph-bootstrap intervals beneath. All contrasts use fine-training seed 0 and the same 2,000 joint resamples of 64 photographs; intervals are marginal, not simultaneous.

Weighting setting	$I(w)$ (pp)	$D(w)$ (pp)
$16 \times 16,\ \alpha = 1/2$ (reference)	+20.70 [16.21, 25.00]	—
$16 \times 16,\ \alpha = 1/4$	+0.05 [−3.37, 3.66]	−20.65 [−24.90, −16.31]
$16 \times 16,\ \alpha = 3/4$	+17.38 [13.96, 20.85]	−3.32 [−6.74, 0.10]
$12 \times 12,\ \alpha = 1/2$	+22.51 [18.85, 26.32]	+1.81 [−2.44, 5.86]
$20 \times 20,\ \alpha = 1/2$	+7.47 [3.32, 11.57]	−13.23 [−17.77, −8.45]
$w_T = 8,\ w_F = 1$	−0.59 [−4.20, 3.22]	−21.29 [−25.68, −16.80]
$w_T = 1,\ w_F = 8$	+3.47 [0.19, 6.74]	−17.24 [−21.63, −12.79]

The reference interaction is +20.70 points (95% CI [16.21, 25.00]). Lowering α to $1/4$ or enlarging the mask to 20^2 reduces this interaction: both corresponding D intervals exclude zero. This inference uses direct paired differences, rather than overlap between separate intervals. The larger mask retains a positive I; the weaker weighting does not establish a support difference.

J.2 REMOVING THE COMPETING-FRAME REQUIREMENT

Let E_c be a correct VGG/SVM prediction at the requested crop, s_ℓ the frame score at that position, τ the original frozen threshold, and $\ell^\star = \arg\max_j s_j$. The alternative event and its difference from original joint success are

$$\begin{aligned} J_{\text{req}} &= \Pr(E_c \cap \{s_\ell > \tau\}), \\ P_{\text{comp}} &= J_{\text{req}} - J = \Pr(E_c \cap \{s_\ell > \tau\} \cap \{\ell^\star \neq \ell\}). \end{aligned} \tag{14}$$

Thus P_{comp} counts class-correct requests excluded solely by a competing frame candidate. The calculation uses all 8,192 stored outputs of the four reference models, preserving the requested crop, threshold, and denominators. The original-event intervals in Table 14 use the common bootstrap draws above, so they can differ slightly from the earlier separately seeded marginal intervals.

Without the competition requirement, the concentrated effect is -2.64 points ($[-5.66, 0.39]$): its point estimate remains negative, but the interval includes zero. The balanced effect remains positive, as does the direct support interaction, $+17.09$ points ($[13.04, 21.14]$). Support dependence therefore survives this sensitivity check, while the evidence for opposite effect signs depends on the original localization definition.

Both events remain frame/VGG proxies. For either support $s \in \{B, S\}$, write $J_{s,w} = J^*_{s,w} + b_{s,w}$, where J^* is the unobserved rate of correct requested texture and placement and b is the mean measurement error. Then the observed interaction contains the additive term $b_{B,w} - b_{B,U} - b_{S,w} + b_{S,U}$. A frozen decision rule does not ensure that these terms cancel. This sensitivity analysis supplies no independent texture labels or human judgments and cannot resolve that bias.

Table 14: **Sensitivity to frame competition.** J_{req} requires a correct class prediction at the requested crop and a passing frame score at that position; it omits the competing-frame argmax requirement. The penalty $P_{\text{comp}} = J_{\text{req}} - J$ counts class-correct requests whose requested frame passes but loses the argmax. The four seed-0 models, the 2,048 requests and all denominators are unchanged, so the two rules differ only in what they credit. Both remain frame/VGG proxies.

Support / contrast	Original J	J_{req}	P_{comp}
Observed counts out of 2,048 requests per model			
Concentrated, uniform	648	672	24
Concentrated, ROI	486	618	132
Balanced, uniform	555	584	29
Balanced, ROI	817	880	63
ROI − uniform effects and their support interaction (pp; 95% intervals)			
Concentrated effect	-7.91 $[-11.08, -4.74]$	-2.64 $[-5.66, 0.39]$	$+5.27$ $[3.52, 7.13]$
Balanced effect	$+12.79$ $[9.72, 15.63]$	$+14.45$ $[11.38, 17.24]$	$+1.66$ $[0.73, 2.78]$
Support interaction	$+20.70$ $[16.21, 25.00]$	$+17.09$ $[13.04, 21.14]$	-3.61 $[-5.62, -1.81]$

K THE POSITION REASSIGNMENT ON A SECOND BACKBONE

The reassignment of Section 5.1 was repeated with SDXL base 1.0 at 512^2 in one seed, under a protocol fixed before training. The four fits keep the B32 support, the seed-0 schedule, the 8,192 updates and 16,384 high-resolution visits, the rank-8 adapter and the zero-initialized coarse branch; routing, the schedule audit and both loss kernels are the SD1.5 modules, and each fit's audit matches its SD1.5 counterpart count for count. Latents are re-encoded with the SDXL VAE from the pixels the SD1.5 latents came from; the SD1.5 VAE applied to those pixels reproduces the SD1.5 latents exactly. The frozen base runs in bfloat16 with a fused NPU attention kernel. Evaluation uses the same coarse samples re-encoded, the same initial noise and the same frame rule on the 128 audited requests, and the frozen class readout on all 2,048.

Table 15: **SDXL at 512^2: both weightings put the extra frames at the training position.** Counts out of the 64 primary requests, and the four-model contrasts with photograph-bootstrap intervals.

Training position	Loss	Extra frame at 0	Extra frame at 1	Requested frame
0	Uniform	36	0	21
0	ROI	57	0	1
1	Uniform	0	53	13
1	ROI	0	57	2

$M(\text{uniform}) = +139.06\ [121.88, 156.25]$, $M(\text{ROI}) = +178.12\ [164.06, 190.62]$
$K = +39.06\ [23.44, 54.69]$; all 128 requests: $+45.31\ [33.59, 55.47]$
K at position 0: $-32.81\ [-45.31, -21.88]$; at position 1: $+6.25\ [-1.56, 15.62]$

By the rule fixed in advance, K replicates: it is positive with an interval excluding zero. The shape differs from SD1.5 in two ways. Uniform weighting already moves the extra frames with the examples, $M = +139.06$ against $+1.56$ on SD1.5, so on this backbone whole-image training inherits the training position as well and ROI weighting adds to it, mainly by emptying the old position; the new one is close to full under both losses. And the models seldom draw the requested frame, least of all under ROI weighting (1 and 2 of 64). The protocol's floor rule, which declares the frame readout uninformative when every arm draws the requested frame in fewer than 16 of 64 requests, is not met, but only because one arm reaches 21. With $M(\text{ROI})$ near its ceiling of $+200$, K is also compressed from above.

Scored over all 2,048 requests with the frozen classifier, the other SD1.5 findings carry over. The four-model interaction on joint success is $+0.10\ [-2.83, 2.78]$, crossing zero as it does in every SD1.5 seed, so here too what follows the training position is unrequested structure rather than success. With concentrated support ROI weighting lowers joint success by $10.21\ [7.32, 13.18]$ and $10.11\ [7.57, 12.70]$ points at the two positions, against 7.91 and 5.76 on SD1.5, and the loss is again in localization rather than class response: detected localization falls by 26.46 and 20.31 points, class response by 2.49 (crossing zero) and 6.69. ALL4 stays between 0 and 4 of 512 in every arm.

Table 16: **SDXL at 512^2, all 2,048 requests.** Counts under the frozen class readout and frame detector; no output was invalid.

Training position	Loss	Joint	Class	Localization	ALL4
0	Uniform	462	736	1094	4
0	ROI	253	685	552	1
1	Uniform	461	816	930	4
1	ROI	254	679	514	0

L WEIGHTING DOSE, AND THE CONSTRUCTION WITHOUT A VISIBLE FRAME

Both follow-ups were fixed before training, use the seed-0 schedule and pair new fits with the published seed-0 models. Their training and sampling loop calls the frozen SD1.5 components; it reproduces the frozen trainer's per-sample losses exactly over the first 16 updates and the frozen sampler's images in 8 of 8, and every new fit's schedule audit matches the reference fit at its position.

Weighting dose. Two fits at each of $\alpha = 1/4$ and $3/4$ complete the reassignment at four doses, $\alpha = \rho$ being uniform weighting (Table 17). The rule fixed in advance, $0 \le K(1/4) \le K(1/2) \le K(3/4)$ on the point estimates, is met, and both new intervals exclude zero; over all 128 audited requests K is $+20.31$, $+67.97$ and $+104.69$. The relocation therefore grows with the share of loss mass on the region, as the coefficient imbalance of Equation 4 does. At $\alpha = 3/4$ the old position empties almost completely, K at position 0 being -79.69 $[-89.06, -70.31]$.

Table 17: **The relocation grows with the weighting.** Extra frames at the two upper positions out of the 64 primary requests, for the fit trained at each position, and the four-model interaction against the seed-0 uniform pair.

Weighting	Trained at 0: at 0 / at 1	Trained at 1: at 0 / at 1	K (pp)
Uniform ($\alpha = \rho$)	2 / 2	1 / 2	0
$\alpha = 1/4$	15 / 2	2 / 11	$+32.81$ $[20.31, 46.88]$
$\alpha = 1/2$ (reference)	30 / 2	1 / 26	$+81.25$ $[64.06, 98.44]$
$\alpha = 3/4$	52 / 4	0 / 30	$+120.31$ $[103.12, 137.50]$

No visible frame. The same 32 photograph–detail pairs were rebuilt without the cyan frame and the gray panel: each texture residual, recovered exactly from the framed target, is added directly to the photograph. It still sums to zero in every 16×16 block, shifted within a block where a pixel would leave $[0, 255]$ (61 of the 64 targets), so each target's coarse observation equals the photograph's own and shows neither the texture nor where it is. The mask comes from the layout geometry, and evaluation starts from block-averaged held-out photographs, because the frozen coarse bank contains frames. The four position arms were trained as in Section 5.1.

The pre-registered estimand, K on high-frequency energy at the two upper panels, is -0.0039 $[-0.0090, 0.0007]$, but the validity rule fixed with it applies: neither ROI fit puts more high-frequency energy at the requested panel than at the others (-0.0007 $[-0.0016, 0.0002]$ and $+0.0005$ $[-0.0009, 0.0018]$), so by that rule the result is uninformative. Descriptively, and not as a test fixed in advance, all four fits draw the texture where their training examples were, whatever the request (Figure 16): the high-frequency difference between the two upper panels is -0.011 under both losses when the examples sat at position 0 and $+0.013$ and $+0.009$ when they sat at position 1, all four intervals excluding zero. Without the frame the observation no longer shows where the request is, and both losses fall back on the training position; the layout code alone does not move them, as in Section 5.4, and ROI weighting adds nothing measurable here.



Figure 16: **Without a visible frame, every fit draws the texture where its examples were.** One request for the lower-left panel, the first primary request by identifier, under the four no-frame fits, left to right: uniform and ROI trained at position 0 (upper left panel), then uniform and ROI trained at position 1 (upper right panel). Original pixels.

M DISCUSSION AND LIMITATIONS

Local weighting carries over the spatial arrangement of the examples it emphasizes, and the cost of doing so depends on how those examples are organized. Interpreting this requires separating what the reassignment identifies, region observability, and measurement validity.

What the reassignment identifies, and what it does not. Moving the training position relocates the extra frames in three independent SD1.5 seeds, while uniform weighting shows no movement at all; on SDXL uniform weighting moves them as well, and ROI weighting adds to that (Appendix K). Because the intervention preserves the photograph–detail identities, the budget, the optimizer state, and the random streams, the transfer cannot be attributed to more supervision or to a different amount of exposure. Three further pairs of fits separate the four things reassignment moves together (Section 5.4). Under concentrated support moving the layout code does little, holding it fixed while the rest moves reproduces the effect in full, and moving the weighted region alone suppresses the extra frames at the old position instead of installing them at the new one. Repeating both cells in two seeds where the weighting is decisive keeps the mask result and not the other: moving the mask removes the gain along with the extra structure in both seeds, while moving the layout code raises the extra structure in one seed and not in the other, and in both costs part of the gain, but only under the weighting. The weighted region therefore governs whether that structure survives in all four fits that move it alone, while the layout code carries the arrangement consistently under neither support. What stays unseparated is the bundle of target, observation and mask, three properties of one training row; because they are, each cell is a cue conflict rather than a clean isolation, and separating them would need independent texture and frame annotations, a different dataset. Weighting the frame is not the mechanism either (Section 5.3), which answers the reading that the effect is an artefact of a region that is at once locator, weighted area and counted object: the second role can be removed without removing the effect, while the other two cannot be removed here, since the frame remains both the cue the observation localizes from and what the geometric rule counts.

What the interactions establish. Balanced support yields gains at all three tested ROI sizes, whereas concentrated support changes sign with both weighting strength and loss coverage. This establishes configuration sensitivity within the measured matrix; it neither makes broad support necessary for improvement nor establishes equivalence among the balanced configurations. The support interaction is positive in three training seeds, but its two halves are not equally stable: balanced gains replicate and the negative concentrated effect does not, so we report the interaction rather than a sign reversal. Those interventions also compensate background weight, and VAE receptive fields mix neighboring regions; resizing the ROI likewise changes its per-pixel weight and the mixture of texture, frame, and background it covers.

From a visible region to an unknown detail. The surviving frame makes our ROI recoverable from the low-resolution observation. The experiment therefore separates locating a relevant region from learning its distinguishing detail. In unmodified photographs, both may be unknown. The tested high-frequency masks address one region-selection rule, but their short continuations do not establish a general solution. Residuals in permitted high-resolution examples offer another source of spatial cues, whose benefit would need testing under the same supervision budget. Removing the frame altogether (Appendix L) leaves the observation with no position cue, and every fit then draws the texture at its training position under either loss; whether local weighting adds to that when a region is unmarked but still locatable is untested, and it is the question that most limits practical reach.

Measurement and transfer. The frozen texture classifier meets its confirmation threshold exactly on constructed held-out examples. Independent assessment of actual generated textures remains unresolved: two automatic alternatives were prepared and neither passed its fixed usability check before any generated output was scored, and their failures are retained rather than replaced by a third attempt. Errors may differ across support and loss configurations, so the confirmation result does not rule out readout bias contributing to the measured interaction. Removing the frame-competition requirement preserves a positive interaction in two seeds but not the third, and makes the concentrated effect's interval cross zero (Appendix J). This distinction matters: a competing frame can cause a joint failure even when the requested crop receives the correct class label. The frame-relocation result is less exposed to this bias, since it counts frame pixels under a geometric rule that is independent of

the scoring argmax, but non-detection is not absence and the rule certifies nothing about texture. We attempted exactly that and can now say what it costs. Five raters judged 1,024 blinded 112×112 crops, frame excluded by construction, under a pre-registered usability gate. The first attempt failed, and diagnosably so: its controls had been drawn through DTD's multi-label annotation to calibrate the classifier, so 31 of 32 positives came from a different directory than the class they stood for and two directories served as both positive and negative. Rebuilding the controls from the four directories themselves recovered most of the deficit, taking one rater from 17 of 32 to 23 and from 0 of 8 to 6 of 8 on the hardest class. It did not clear the bar. Three raters reached 62 of 96 positives, 64.6%, well above the 25% chance rate and below the 75% this classifier is held to and achieves on the same kind of material, so on its own domain these raters are weaker than the instrument they were meant to check. Under the pre-registered rule none passed, and all 1,024 ratings were discarded in full rather than the threshold being moved. Agreement separates the two regimes: 66.0% pairwise on real textures against 24.3% on generated crops, unanimous on 54.2% against 1.5%, which says the generated crops are markedly more ambiguous than the textures they are built from. Three untrained raters are a convenience sample and this does not show that no human could do it; it does bound what an untrained human panel delivers here, and it is why the readouts in this paper are classifier and geometry based rather than human.

Native 4K separates the two findings. The completed SDXL four-model set gives a support interaction of -3.65 points $[-8.85, 1.56]$ on the matched positions and -2.73 $[-7.42, 2.34]$ on all requests: the interval crosses zero and the point estimate runs opposite to the 512×512 result, so the support interaction does not replicate at this resolution and backbone. The side effect does carry over: detected detail in unrequested regions rises under ROI weighting in both supports, by $+13.54$ points $[6.77, 20.31]$ and $+8.85$ $[3.65, 14.06]$, measured by a different backbone, resolution and detector. That is consistency of the side effect rather than replication of the inheritance result, because the reassignment intervention has not been repeated at this resolution and the 4K detector counts template responses rather than cyan frames. All evaluated 4K models have zero successful four-request quartets, and the constructed residuals, opaque request codes and low 4K detail response leave generalization to natural detail and reliable control open.

N FORMAL PROPERTIES OF THE CONSTRUCTION AND MEASUREMENTS

The statements below establish properties of the observation, loss coefficients, and scores. They do not predict an optimization trajectory or prove the sign of the empirical ROI effect.

Proposition 1 (observation ambiguity). Vectorize an image and let A be its block-averaging matrix, applied independently to each channel. Its nonempty, disjoint blocks give full row rank. The orthogonal projector onto its nullspace is

$$P_A = I - A^\top (AA^\top)^{-1} A, \qquad AP_A = 0, \qquad P_A^2 = P_A. \tag{15}$$

Consequently, every residual $r \in \ker A$ leaves the observation unchanged. For the counterfactual distribution that draws D uniformly from the four detail classes at each background $B = b$ and layout ℓ, the constructed observation satisfies

$$L_{b\ell d} = L_{b\ell}, \qquad I(D; L \mid B, \ell) = 0, \qquad H(D \mid L, B, \ell) = H(D \mid B, \ell) = 2 \text{ bits}. \tag{16}$$

Proof. Direct multiplication gives $AP_A = 0$ and $P_A^2 = P_A$; P_A is symmetric because $(AA^\top)^{-1}$ is symmetric. In the implemented construction, each residual instead has an exact integer zero sum over the intersection of a block with the texture interior, and is zero outside the interior. Its sum over the complete block is therefore zero, including boundary blocks, so $Ar = 0$. This establishes the same nullspace property without identifying the integer procedure with an orthogonal projection. Finally, L is constant across D conditional on (B, ℓ), proving the information identity. The uniform counterfactual distribution is not the sparse training-support distribution. The generator also receives a detail request, so the identity imposes no one-in-four ceiling on conditional generation. It asserts no nullspace property of the VAE.

Proposition 2 (spatial coefficient marginal). Let Ω contain N latent positions, and let S contain K equally visited supervised conditions with nonempty masks $M_i \subsetneq \Omega$. For $\alpha \in [0, 1]$, define

$$\begin{aligned} w_{i,\alpha}(u) &= \frac{\alpha\, \mathbf{1}\{u \in M_i\}}{|M_i|} + \frac{(1-\alpha)\, \mathbf{1}\{u \notin M_i\}}{N - |M_i|}, \\ \mu_{S,\alpha}(u) &= \frac{1}{K} \sum_{i=1}^{K} w_{i,\alpha}(u), \qquad \sum_{u \in \Omega} \mu_{S,\alpha}(u) = 1. \end{aligned} \tag{17}$$

Assume each mask equals its assigned geometry among m disjoint masks $M^{(j)}$ of common area ρN, with $0 < \rho < 1/m$. Let p_j be the supervised-visit fraction assigned to position j. Define $c_{\text{out}} = (1-\alpha)/(1-\rho)$ and $\kappa = \alpha/\rho - c_{\text{out}} = (\alpha - \rho)/[\rho(1-\rho)]$. Then

$$N\mu_{S,\alpha}(u) = \begin{cases} c_{\text{out}} + \kappa p_j, & u \in M^{(j)}, \\ c_{\text{out}}, & u \notin \bigcup_{j=0}^{m-1} M^{(j)}. \end{cases} \tag{18}$$

Writing $q_j = N\mu_{S,\alpha}(u)$ for $u \in M^{(j)}$ gives the centered relation

$$\mathbf{q} - \left(c_{\text{out}} + \frac{\kappa}{m}\right) \mathbf{1} = \kappa \left(\mathbf{p} - \frac{1}{m}\mathbf{1}\right). \tag{19}$$

When $\alpha = \rho$, $c_{\text{out}} = 1$ and $\kappa = 0$: weighting is exactly uniform for every support. For $\rho < \alpha < 1$, the largest-to-smallest coefficient ratio is finite and equals

$$\begin{aligned} R(S) &= \frac{\max_u \mu_{S,\alpha}(u)}{\min_u \mu_{S,\alpha}(u)} = 1 + \frac{\kappa}{c_{\text{out}}} \max_j p_j, \\ R_{\text{equal spread}} &= 1 + \frac{\kappa}{m c_{\text{out}}}. \end{aligned} \tag{20}$$

At fixed θ, define $\bar{e}_i(u) = \mathbb{E}[e_\theta(u) \mid i]$, where the expectation averages over timestep, noise, and caption dropout. Drawing condition i uniformly from S gives the objective over supervised visits:

$$\begin{aligned} \mathcal{J}_H &= \sum_{u \in \Omega} \left[\mu_{S,\alpha}(u) \mathbb{E}_i \bar{e}_i(u) + \mathrm{Cov}_i(w_{i,\alpha}(u), \bar{e}_i(u))\right], \\ \mathcal{J}_{\text{scheduled}} &= h_H\, \mathcal{J}_H. \end{aligned} \tag{21}$$

Here h_H is the fraction of scheduled slots with supervision; the remaining slots contribute zero loss.

Proof. Each w_i sums to $\alpha + (1 - \alpha) = 1$. At position j, a fraction p_j of visits assigns coefficient $\alpha/(\rho N)$ and the remainder assigns c_{out}/N, giving Equation 18. Centering uses $\sum_j p_j = 1$. For $\rho < \alpha < 1$, $\kappa > 0$ and $c_{\text{out}} > 0$; the nonempty region outside all masks attains the minimum, yielding $R(S)$. At $\alpha = 1$ this minimum is zero, so the finite-ratio formula does not apply. Finally, apply the covariance identity $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + \text{Cov}(X, Y)$ at each pixel and count supervised slots.

Experimental instantiation. With $\rho = 1/16$, $\alpha = 1/2$, and $m = 4$, the per-condition weight inside the ROI is 8 relative to uniform, while $c_{\text{out}} = 8/15$ and $\kappa = 112/15$. Equal spread has the average relative coefficient $q_j = 12/5$; the concentrated and equal-spread ratios are 15 and $9/2$. The scheduled fraction is $h_H = 16{,}384/(8{,}192 \times 16) = 1/8$. Initial and balanced spread share μ, but this does not imply equal expected loss: the conditional-error and mask–error covariance terms can both differ.

Proposition 3 (joint and quartet identities). Let E_c denote a correct class response and E_ℓ a correct detected location on the same request population. Write $C_{\text{cls}} = \Pr(E_c)$ and $W = \Pr(E_c \cap E_\ell^c)$. For candidate-minus-reference differences,

$$J = C_{\text{cls}} - W, \qquad \Delta J = \Delta C_{\text{cls}} - \Delta W. \tag{22}$$

For G complete, equally weighted quartets, let $y_{gd} \in \{0, 1\}$ denote joint success for detail d. Then

$$J = \frac{1}{4G} \sum_{g=1}^{G} \sum_{d=1}^{4} y_{gd}, \qquad A_4 = \frac{1}{G} \sum_{g=1}^{G} \prod_{d=1}^{4} y_{gd}, \qquad \max(0, 4J - 3) \le A_4 \le J. \tag{23}$$

Proof. Partition E_c by E_ℓ to obtain the first identity, then subtract the uniform identity from the ROI identity. Within each quartet, $\sum_d y_{gd} - 3 \le \prod_d y_{gd} \le \frac{1}{4} \sum_d y_{gd}$. Average and use $A_4 \ge 0$. Neither statement assumes independence; neither establishes image-level transitions or a causal mechanism. In particular, joint success below 75% gives no positive lower bound on four-way success.