UNSURE WHAT EXISTS, NOT MISSING PHYSICS: INTERMEDIATE DENOISING ESTIMATES VIOLATE SUPPORT CONSTRAINTS IN 3D SCENE DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intermediate denoising estimates of 3D scene diffusion models are often read as evidence about learned physical constraints: a lifted object is pulled back to the floor, or a large residual gap is reported as a violation. We show that, in joint semantic–geometric diffusion, the second reading is usually wrong for a simple reason. The clean-geometry estimate is a posterior mean over what each slot might contain, and for an affine constraint such as floor support this mean leaves the constraint plane exactly when the posterior places mass on states that do not satisfy it: an empty slot, whose zero-padded geometry decodes to a floating box, or a non-floor category. We test this account with controlled semantic-token interventions on official test rooms of three room-specific MiDiffusion checkpoints and, with continuous class channels, on DiffuScene. A predictor with no fitted parameter, the network's own auxiliary occupancy posterior times the decoded empty coordinate plus class-conditional training means, reproduces the geometry head's floor gap (R^2 up to 0.96), tracks its growth with noise level, and matches the sign and magnitude of every context intervention. Replaying complete reverse trajectories step by step shows that the estimate settles on the floor before the categorical token commits, that tokens commit exactly once, and that every final failure among 592 known-scene trajectories is a target whose occupancy was unresolved at the intervention step. The consequence is practical: projecting intermediate estimates onto the floor at every step without conditioning on occupancy adds 0.8–1.1 spurious objects per scene and raises the colliding-object fraction by 24–35 points, and occupancy-blind collision guidance on the noisy state likewise creates objects; gating on the occupancy argmax removes most of these objects but does not make collision guidance effective. Paired response scores cannot see any of this, because they cancel the shared offset. Constraint diagnostics and guidance on intermediate estimates must therefore be conditioned on occupancy.

1 INTRODUCTION

Object-based scene generators predict both what occupies a room and where each object is placed. Diffusion models such as DiffuScene and MiDiffusion learn these layouts from furnished-room data (Fu et al., 2021; Paschalidou et al., 2021; Tang et al., 2024; Hu et al., 2026), and their intermediate denoising estimates are an appealing window onto what they learned: lift a chair above the floor, and the single-step clean-geometry estimate pulls it back down. Such responses have been used to argue that a support constraint is encoded, and intermediate estimates are also the quantity that physics-aware guidance methods edit or penalise during sampling (Yang et al., 2024; Wu et al., 2026). Both uses presuppose that a violation in the intermediate estimate is of the same kind as one in a finished scene.

This paper shows that in joint semantic–geometric diffusion the presupposition fails, and identifies the mechanism. A slot of the intermediate state may still become a chair, a ceiling lamp, or nothing at all, and the clean-geometry estimate is a mean over these possibilities. For an affine constraint like floor support $g = y - h - f = 0$, a mean of floor-supported states stays on the plane however they are weighted (§2); it leaves the plane only through states that are not on it. Two such states exist by construction: the empty slot, whose zero-padded geometry decodes to a box floating 80–94 cm

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.

above the floor, and the non-floor categories. Intermediate support violations are therefore a readout of occupancy uncertainty, not of missing physical knowledge. We establish this in four steps, each with controlled interventions on official test rooms and reviewed, replayable runs.

- **Single-step estimates are occupancy-weighted class means (§4).** Restoring a masked target token reduces the original-slot gap error from 13–17 cm to about 2 cm in three room-specific MiDiffusion checkpoints while every paired ranking stays perfect. A predictor with no fitted parameter, the auxiliary occupancy posterior times the decoded empty coordinate plus class-conditional training means, reproduces the geometry head (R^2 0.96 and 0.91 under native corruption), tracks the rise of the error with noise level, and matches the sign and magnitude of every context intervention, including the previously unexplained effect of padding tokens: each masked padding slot raises the target's empty posterior by 0.05–0.06, a masked real object does not.
- **The sampler resolves occupancy before geometry (§5).** A bit-exact replay of 720 reverse trajectories shows that the clean estimate of a masked target settles within 5 cm of the floor at a median step of about 200, before its token is sampled at a median step of 127–174, and that tokens commit exactly once. Every retention or category failure among 592 known-scene trajectories is a target that was masked at the intervention step, and the empty posterior at that step predicts which targets will be deleted.
- **Occupancy-blind constraints on intermediate estimates hallucinate objects (§6).** Projecting every slot's estimate onto the floor at every step adds 0.8–1.1 objects per scene and raises the colliding-object fraction by 24–35 points; gating on the occupancy argmax removes these side effects, and a one-time projection has no effect because the posterior mean weights the edited estimate by only 0.006. Box separation inside the posterior mean does not detectably change collisions or object counts; in the noisy state it creates objects without gating and, gated, does not reliably reduce collisions.
- **The mechanism, not its discrete symptoms, transfers across architectures (§7).** On DiffuScene, whose class one-hot and shape latent are diffused continuously, removing the semantic signal or asserting the empty class moves the estimate off the floor in the same order with ranking unchanged, whereas two phenomena that depend on absorbing-state masking, the near-equivalence of wrong floor categories and the padding-context effect, do not transfer.

Paired response scores cancel the shared offset and pass under all of these conditions, including projection onto the wrong plane (§2). The argument needs only an affine coordinate constraint and an intermediate estimate that averages over discrete alternatives; padded element sets in document, sketch, and floorplan generators share these properties, and §9 states the expected scope.

2 Affine constraints under occupancy uncertainty

Setting. Denoising models combine a learned predictor with a specified reverse process (Ho et al., 2020; Song et al., 2021b;a; Karras et al., 2022). A joint scene diffusion model represents a room as N slots. Slot i carries continuous geometry z_i (translation, half-extents, yaw) and a categorical state c_i drawn from the real classes, an `<empty>` class for unoccupied slots, and, in the absorbing discrete process of MiDiffusion, a `<mask>` state (Austin et al., 2021; Hu et al., 2026). Training encodes real geometry into $[-1, 1]$ and pads absent slots with zeros. At reverse step t the denoiser receives the noised geometry x_t and the current tokens c_t and returns a clean-geometry estimate $\widehat{x}_0$ and, in the mixed model, auxiliary clean-class logits. For an upright box with vertical centre y, half-height h, and floor height f, the signed floor gap is the affine function

$$g(z) = y - h - f. \tag{1}$$

Proposition (affine mixtures). Let z have any integrable conditional distribution given the noisy state, for instance a mixture over the categorical states the slot may take. Because g is affine,

$$g(\mathbb{E}[z \mid x_t, c_t]) = \mathbb{E}[g(z) \mid x_t, c_t]. \tag{2}$$

If every state with positive weight satisfies $g = 0$, the conditional mean does too, whatever the weights. Conversely the mean can leave the plane only through states that are not on it. With $p_\varnothing$ the posterior weight of the empty state, $\mu_\varnothing = g(D(0))$ the decoded gap of zero-padded geometry, and μ_c the mean gap of class c,

$$\mathbb{E}[g(z) \mid x_t, c_t] = p_\varnothing\, \mu_\varnothing + \sum_c p_c\, \mu_c, \tag{3}$$

which is zero for floor-supported classes and large for the empty state (93.99 cm in the bedroom encoding, 80.26 cm in the living room) and for lamps (training means 179–234 cm). Equation 3 is a statement about any posterior-mean estimator; it does not assert that the network computes a mixture. Section 4 tests whether the network's own auxiliary posterior, inserted for p, reproduces its geometry head.

Three predictions follow. (i) Intermediate gap error under semantic uncertainty is governed by $p_\varnothing$ and by non-floor mass, not by the identity of the floor-supported class; asserting any floor class should remove most of it, asserting a lamp should produce a large but legitimate lamp-height coordinate, and asserting <empty> should produce $\mu_\varnothing$. (ii) Because the geometry itself becomes informative about occupancy as noise decreases, $p_\varnothing$ under a fixed masked token should fall with t, and the joint sampler should be able to place a slot on the floor before its token is sampled. (iii) Final samples fail the constraint only where occupancy or category is resolved wrongly, and enforcing the constraint on an intermediate estimate without conditioning on occupancy should move empty slots off their zero prior and create objects.

Why paired response scores cannot see this. The support-response diagnostics that motivated this work compare a lifted input I with a support-preserving control C that shares the same noise and tokens. With g_j and $\hat{g}_j$ the actual encoded-and-decoded input and predicted gaps and s the sign of the nominal lift, the implemented score is $S_j = s(g_j - \hat{g}_j)$ and

$$S_I - S_C = s\big[(g_I - g_C) - (\hat{g}_I - \hat{g}_C)\big]. \tag{4}$$

Any output offset shared by the pair cancels. Two analytic counterexamples on the exploratory 600-pair construction make the consequence concrete (Table 10, Appendix C): projection onto a plane 20 cm above the floor and a constant vertical output both rank every pair correctly. $\mathrm{PairAcc}$, the fraction of strict $S_I > S_C$ comparisons, is exactly 1 in all 19 semantic conditions of §4 while the absolute readout ranges from 2 to 94 cm. We therefore report absolute original-slot gaps, the occupancy posterior, and final joint events, and use $\mathrm{PairAcc}$ only as a check that the perturbation direction is recovered.

Final joint events. For a target slot we record retention R of the original slot in the decoded sample, agreement C of its sampled category with the reference, membership A of that category in the predefined floor-applicable set, and H: positive raw half-extents with $|g| \leq 5$ cm. Supported occupancy is $S = A \wedge H$ and joint success is $J = C \wedge A \wedge H$. Deletion is never counted as repair, and C measures slot and class agreement, not mesh identity.

3 Experimental protocol

Models and cohorts. We use the released floor-conditioned bedroom, living-room, and dining-room checkpoints of MiDiffusion (Hu et al., 2026), trained on 3D-FRONT (Fu et al., 2021), and the released unconditional bedroom and living-room checkpoints of DiffuScene (Tang et al., 2024). The confirmation cohorts are official room-level test scenes with available assets, excluding known source overlap with the released training splits and with the exploratory cohort: 107 bedroom scenes with 309 targets, 189 living-room scenes with 567 targets, and 175 dining-room scenes with 524 targets. Source mapping is incomplete, so we claim room-level, not house-level, independence.

Perturbations, semantics, and readouts. Targets are floor-supported objects with a reference gap within 1 cm, chosen by category coverage and a fixed hash before any prediction. Each target receives four Gaussian draws and, for MiDiffusion, four categorical draws from the model's forward process, crossed with vertical lifts of 5, 10, and 20 cm; the control moves centre and half-height equally so that the gap is preserved. The primary $t = 300$ matrix has 19 semantic conditions: the correct category, four incorrect floor-supported categories selected from training-only size statistics, <mask>, and <empty>, each under clean or native context, plus restoration of real-object or padding context and masking of one real or padding slot. For DiffuScene the same names denote edits of the class channels of the noised state (Appendix A). The primary readouts are the original-slot gap of the single-step clean estimate, the rate above 20 cm, $\mathrm{PairAcc}$, and the auxiliary occupancy posterior; §5 and §6 add full reverse trajectories and final joint events.

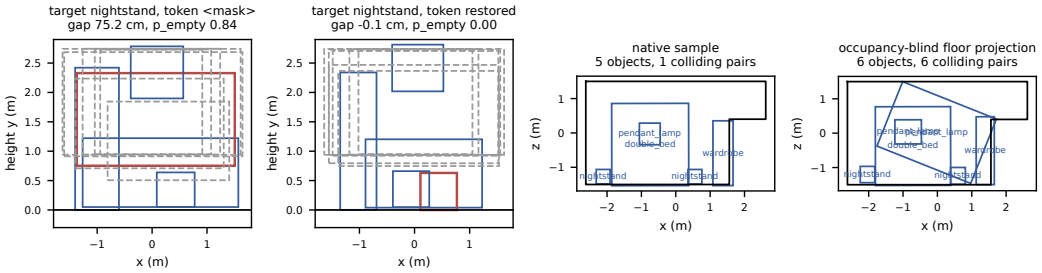


Figure 1: Two hash-selected bedroom examples. Left: side view of the $t = 300$ clean estimate for a natively masked target (red) under the <mask> token and after its category is restored; dashed boxes are clean estimates of reference-empty slots (readouts, not objects). Right: final samples of one native-generation trajectory under native continuation and under occupancy-blind every-step floor projection, which turns a padding slot into a sixth object overlapping the furniture (§6).

Estimation, verification, and disclosure. Effects are scene-equal means of paired case differences with 2,000 source-cluster bootstrap resamples for pointwise 95% intervals; repeated draws and lifts are not independent scenes. Category comparisons use prespecified bounds of ± 1 cm and ± 5 percentage points; a non-significant contrast is never read as equivalence. Every run passes input-edit contracts, explicit no-ops, and batch-size equivalence within 2×10^{-5} in model coordinates; trajectory replays reproduce the saved samples exactly. Protocols were frozen before their outputs; Appendix A records the design changes made after earlier results had been seen.

4 Single-step estimates are occupancy-weighted class means

Restoring the token removes the error; ranking never changes. All 19 conditions of the primary $t = 300$ matrix have exactly $\mathrm{PairAcc} = 1$ in every checkpoint (Table 4, Appendix B). Restoring the masked target's category lowers the scene-equal original-slot gap error from 17.35 to 2.25 cm in bedrooms, from 13.84 to 2.40 cm in living rooms, and from 13.48 to 2.24 cm in dining rooms; the paired effects are -15.10 cm $[-17.06, -13.35]$, -11.44 cm $[-12.51, -10.44]$, and -11.24 cm $[-12.44, -10.17]$. Forcing every target to <mask> gives 50.6, 39.0, and 38.9 cm; forcing <empty> gives 93.5, 80.5, and 81.4 cm, which are the decoded zero-padding coordinates of each encoding rather than floating furniture. Lamp tokens give 170–229 cm, the training heights of those classes. Figure 1 (left) shows one hash-selected case. Each of the four incorrect floor-supported categories retains most of the recovery (2.3–3.5 cm), but the class matters at the centimetre level: the prespecified four-category average exceeds the correct token by 0.63 cm $[0.44, 0.88]$, 0.73 cm $[0.63, 0.83]$, and 0.94 cm $[0.79, 1.09]$, and the dining-room value fails the ± 1 cm bound we fixed in advance.

The network's own occupancy posterior reproduces its geometry head. We evaluate Equation 3 with p taken from the softmax of the auxiliary clean-class logits at the target slot, $\mu_\varnothing$ from the encoding bounds, and μ_c from the official training scenes; the predictor has no fitted parameter. Table 1 and Figure 2a compare it with the geometry head case by case. Under native corruption the predictor explains 96% and 91% of the variance of the predicted gap (RMSE 6.1 and 7.4 cm against 30.2 and 24.3 cm for a constant predictor); under a forced mask it explains 90% and 81%, and across the four context interventions 76–90%. The empty term alone explains far less (R^2 0.67 and 0.61 native), so the non-floor classes contribute as well. With the token held at <mask> and the context fixed to its $t = 300$ realisation, the empty posterior rises monotonically with the noise level, 0.03, 0.15, 0.26, 0.40 in bedrooms from $t = 100$ to 300, in step with the gap error, 8.5, 23.5, 36.3, 50.6 cm, and the case-level correlation exceeds 0.94 at every level (Figure 2d).

Context interventions are changes in the occupancy posterior. The same predictor reproduces the sign and magnitude of the five context interventions and of target restoration (Figure 2c; scene-equal changes of the signed gap); in two living-room contrasts the predicted change is 2–3 cm smaller than the observed one with non-overlapping intervals, so the agreement is quantitative but not exact. Restoring only the padding tokens around a masked target changes the geometry head by -13.4 cm

Table 1: Parameter-free semantic-head predictor $p_\varnothing \mu_\varnothing + \sum_c p_c \mu_c$ versus the geometry head at $t = 300$ (lifted members, pooled cases, cm; RMSE intervals are source-cluster bootstrap; "mask" is a forced <mask> target; "const." is the RMSE of the within-condition mean predictor).

Room	Condition	n	$p_\varnothing$	RMSE [95% CI]	R^2	R^2 empty	Const.
Bedroom	native	14,832	0.12	6.06 [5.42, 6.74]	0.960	0.669	30.17
	native, masked subset	4,584	0.39	9.94 [8.74, 11.11]	0.911	0.140	33.36
	mask, native ctx.	14,832	0.40	10.15 [9.02, 11.29]	0.897	0.192	31.66
	mask, clean ctx.	14,832	0.20	13.03 [10.40, 15.69]	0.887	0.339	38.71
Living room	native	27,216	0.10	7.43 [6.81, 8.05]	0.907	0.614	24.35
	native, masked subset	8,448	0.32	12.36 [11.21, 13.59]	0.818	0.149	28.93
	mask, native ctx.	27,216	0.32	12.53 [11.58, 13.51]	0.806	0.074	28.47
	mask, clean ctx.	27,216	0.16	17.14 [14.85, 19.51]	0.731	0.352	33.05

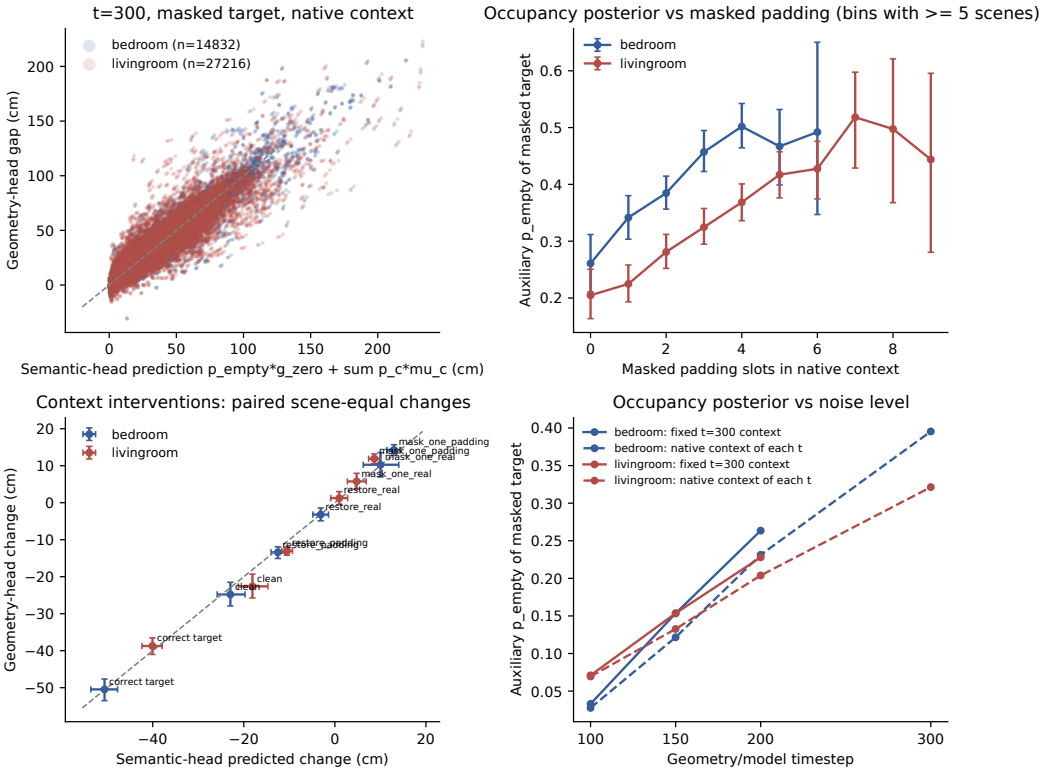


Figure 2: The auxiliary occupancy posterior explains the geometry head. (a) Per-case semantic-head prediction versus predicted floor gap for masked targets at $t = 300$. (b) Empty posterior of a masked target versus the number of masked padding slots in its context. (c) Geometry-head change versus predicted change for the five context interventions and target restoration, scene-equal with 95% cluster-bootstrap intervals. (d) Empty posterior of a masked target versus noise level. Agreement between two outputs of one network is consistency, not causal mediation.

$[-15.1, -11.9]$ and the predictor by -12.5 cm $[-14.0, -10.9]$ in bedrooms, with the empty posterior falling by 0.15; restoring only real-object tokens changes the geometry head by -3.2 cm and the empty posterior by only $+0.02$. Masking one additional padding slot from a clean context raises the head by $+14.3$ cm and the predictor by $+13.0$ cm. A within-scene regression of the target's empty posterior on the number of masked context slots gives $+0.060$ $[0.052, 0.066]$ per masked padding slot and -0.012 $[-0.024, -0.001]$ per masked real object in bedrooms, and $+0.048$ $[0.043, 0.052]$ versus -0.019 $[-0.024, -0.014]$ in living rooms. Masked slots are exchangeable to the network: the more masked slots that are in fact empty, the more likely any masked slot is empty. Padding tokens matter because they carry occupancy information, not because they are scene semantics.

Table 2: Final events of known-scene trajectories by the native $t = 300$ token of the target (scene-equal %, 95% cluster-bootstrap intervals for J). Every native-branch failure lies in the masked stratum. Under native continuation a masked target ends deleted in 32.1% / 26.8% of trajectories, as another floor class in 34.9% / 48.8%, as a non-floor class in 2.8% / 7.1%, and as the reference class in 30.2% / 17.3% (bedroom / living room).

Room	Stratum, branch	n	R	C	A	J [95% CI]
Bedroom	known, native	156	100	100	100	100 [100, 100]
	masked, native	58	67.9	30.2	65.1	30.2 [18.7, 42.3]
	masked, target restored once	58	100	100	100	100 [100, 100]
	masked, context restored once	58	80.2	53.8	73.6	53.8 [40.6, 67.0]
Living room	known, native	280	100	100	100	100 [100, 100]
	masked, native	98	73.2	17.3	66.1	17.3 [10.1, 25.0]
	masked, target restored once	98	100	100	100	100 [100, 100]
	masked, context restored once	98	86.3	56.5	83.9	56.5 [46.4, 66.7]

What this does and does not establish. The interventions establish causal dependence of the single-step estimate on the semantic input, and the parameter-free agreement establishes quantitative consistency with an occupancy-weighted class mean; neither identifies an internal mixture computation, and the residual class dependence shows that category identity is not irrelevant.

5 THE SAMPLER RESOLVES OCCUPANCY BEFORE GEOMETRY

Design. From the same $t = 300$ state we run the unmodified joint sampler to completion under four branches with shared future randomness: native continuation, one-time restoration of the target token, one-time restoration of all other tokens, and an explicit no-op. Known-scene trajectories use one preselected target per confirmation scene with a 10 cm lift and two draws (214 bedroom, 378 living-room trajectories); a native-generation cohort uses 32 floorplans and two seeds per room with the baseline's final class as a retrospective reference. No token is clamped after the intervention. We stratify every trajectory by whether the target token was <mask> at $t = 300$ in the native draw, and we replay all 301 steps with a step-by-step replica of the sampler that reproduces the saved final slots and snapshots exactly (maximum deviation 0), recording each slot's input and output token, clean-estimate gap, and auxiliary empty posterior at every step.

Every final failure is an unresolved-occupancy target. Table 2 gives the final events by stratum. Targets whose token was known at $t = 300$ are retained with the reference category and supported within 5 cm in 156 of 156 bedroom and 280 of 280 living-room trajectories. All 39 bedroom and 80 living-room joint failures of the native branch occur in the masked stratum, where joint success is 30% $[19, 42]$ and 17% $[10, 25]$. A masked target ends deleted in 32% and 27% of trajectories, as a different floor-supported category in 35% and 49%, as a non-floor category in 3% and 7%, and as the reference category in only 30% and 17%. Whatever floor-supported object finally occupies the slot is supported: in no branch of any cohort does a retained floor-applicable target miss the 5 cm threshold, so $S = A$ and $J = C$ case by case. One-time restoration of the target token raises the masked stratum to 100% on every event because it resolves identity, not because it repairs geometry: on the targets retained and floor-applicable in both branches the final gap changes by -0.003 cm $[-0.010, 0.006]$ and $+0.007$ cm $[-0.001, 0.016]$. The saved intermediate readouts (Figure 4, Appendix B) descend from 2 cm to 0.2 cm in the known stratum and stay at 32–45 cm through $t = 0$ in the masked stratum, because a third of its slots become empty or change class.

Geometry settles before the token commits; tokens commit once. In the step-resolved replay of masked targets under native continuation (Table 8, Appendix B), the clean estimate of a retained target enters and stays within 5 cm of the floor at a median step of 176–210, whereas the input token first leaves <mask> at a median step of 127–174 over all masked targets; geometry precedes the token in 28 of 58, 55 of 98, 16 of 26, and 13 of 20 trajectories, and the token precedes geometry in at most one per cohort. Once a token leaves <mask> it never changes again: in 720 native trajectories the output token of a masked target changes exactly once, no trajectory switches between two real classes, and none of the 518 known-stratum targets changes token in 301 steps. The empirical 100%

Table 3: Every-step floor projection of the clean estimate, guided minus native (scene-equal, 95% cluster-bootstrap intervals). Objects: sampled objects per scene; collisions, boundary: colliding and out-of-bounds object fractions in percentage points.

Cohort	Arm	Objects	Collisions (pp)	Boundary (pp)
Bedroom, known (107)	occupancy-blind	$+0.83$ [0.69, 0.98]	$+34.7$ [30.3, 39.0]	$+4.0$ [0.7, 7.3]
	gated	$+0.00$ [-0.01, 0.02]	$+0.1$ [-0.2, 0.6]	$+0.1$ [-0.5, 0.7]
Living room, known (189)	occupancy-blind	$+0.80$ [0.67, 0.92]	$+26.3$ [24.0, 28.7]	-0.5 [-1.8, 0.7]
	gated	$+0.01$ [-0.01, 0.02]	-0.1 [-0.5, 0.3]	$+0.1$ [-0.1, 0.4]
Bedroom, gen. (32)	occupancy-blind	$+1.09$ [0.70, 1.50]	$+34.8$ [27.8, 41.8]	$+2.7$ [-1.9, 7.7]
	gated	-0.02 [-0.05, 0.00]	$+0.9$ [0.0, 2.0]	$+1.3$ [0.1, 2.8]
Living room, gen. (32)	occupancy-blind	$+0.88$ [0.59, 1.17]	$+24.2$ [19.5, 29.0]	$+1.7$ [-0.9, 4.4]
	gated	$+0.00$ [-0.05, 0.05]	$+0.0$ [-1.0, 1.1]	$+0.1$ [-0.7, 0.9]

category rate after a one-time token restoration is therefore a sample property of a sampler that commits once, not a guarantee of the transition kernel. The trajectories that end deleted are visible early: their auxiliary empty posterior at $t = 300$ averages 0.56 in both rooms against 0.19–0.42 for retained targets, their clean estimate never settles within the 5 cm band (it enters it only transiently in 3 of 28 living-room cases), and their output token is `<empty>` for a median of 133–166 of the 301 steps. The intermediate readout carries information about the final occupancy outcome; it does not carry information about physical failure, because no retained floor object fails.

Native generation. The native-generation cohort has a selected-target ceiling: the reference is the baseline's own final class, so native continuation is correct by construction. Two observations survive it: restoring all context tokens at $t = 300$ lowers joint success from 100% to 89% in both rooms with every loss in the masked stratum, and the same ordering holds.

6 Occupancy-blind constraints on intermediate estimates create objects

If intermediate support violations are occupancy readouts, then enforcing the constraint on the intermediate estimate without conditioning on occupancy should move empty slots away from their zero-geometry prior and eventually make the network treat them as occupied. We test this with the simplest intermediate-estimate intervention: at every reverse step $t \leq 300$ the clean-geometry estimate is edited so that selected slots satisfy $g = 0$ (centre height set to half-height), the posterior mean of x_{t-1} is formed from the edited estimate with the native variance and the shared Gaussian draw, and the joint sampler is otherwise unchanged. The *occupancy-blind* arm edits every slot; the *gated* arm edits only slots whose auxiliary clean-class argmax is a floor-applicable real class. Both arms start from the saved $t = 300$ states of §5, and the native branch again reproduces the saved samples exactly. This is a diagnostic of a class of interventions, not a proposed method.

Table 3 reports scene-equal paired differences from the native branch. The occupancy-blind projection adds 0.83 $[0.69, 0.98]$ and 0.80 $[0.67, 0.92]$ sampled objects per known scene and 1.09 and 0.88 per generated scene, raises the colliding-object fraction by 35, 26, 35, and 24 percentage points, and increases boundary violations in bedrooms. In living rooms almost all of the added objects are floor-supported categories. The gated projection leaves object counts and collisions unchanged within their intervals; among the three quantities of Table 3 its one interval excluding zero is a 1.3-point increase of boundary violations in generated bedrooms $[0.1, 2.8]$. Both arms lower the final floor-gap error of retained floor objects by 0.2–0.8 cm, but the $t = 0$ edit fixes the selected slots' geometry by construction, so this is not evidence of guidance quality. In the native-generation cohort both arms lose 1.6–4.7 points of target retention or category agreement, with intervals reaching zero, so persistent geometric editing also perturbs identity sampling slightly.

A one-time projection at $t = 300$ changes nothing (Appendix B): the DDPM posterior mean (Ho et al., 2020) weights the edited estimate by $\beta_t\sqrt{\bar{\alpha}_{t-1}}/(1-\bar{\alpha}_t) \approx 0.006$ at that step, and the difference has decayed to 10^{-4} cm by $t = 0$, whereas token edits persist because the absorbing transition keeps an unmasked token.

Where the edit enters decides whether it persists. We also separate overlapping box footprints at every step from $t = 300$ to $t = 50$, with the last 50 steps native, for all slots or only for slots whose auxiliary argmax is a real class, in three ways: inside the posterior mean as above, carried into the noisy state with the network's noise estimate held fixed, and as the gradient of a penetration loss propagated through the denoiser to the state. Inside the posterior mean, where the edit enters with weight 0.006 at $t = 300$, neither arm detectably changes collisions, object counts, or boundary violations: every interval of Table 6 includes zero. Carried into the state, the occupancy-blind edit creates objects in seven of the eight cohort-by-form cells, 0.23 to 1.09 per scene, and the state form raises the colliding-object fraction by 6–9 points, with three of four intervals excluding zero (Table 7, Appendix B). Gating removes most of the added objects (gated changes of -0.02 to $+0.12$ per scene), but it does not reliably reduce collisions: the gated colliding-object fraction changes by -3.4 to $+2.8$ points and every interval includes zero, so the criterion we fixed in advance for calling gated guidance a collision method is not met. The gated state form also lowers category agreement in generated scenes by 10.9 and 12.5 points. Occupancy gating is therefore needed to keep guidance on intermediate estimates from creating objects, but it is not sufficient.

7 What transfers: the mechanism, not its discrete symptoms

Three room-specific checkpoints. The bedroom, living-room, and dining-room MiDiffusion checkpoints replicate the primary effects (Table 4): target restoration -15.1, -11.4, and -11.2 cm; padding restoration -12.8, -11.8, and -11.3 cm; perfect ranking in all 19 conditions; and empty readouts at the decoded zero-padding coordinate of each encoding. The residual dependence on the exact floor category is heterogeneous: the four-category average exceeds the correct token by 0.6, 0.7, and 0.9 cm, and the dining-room interval $[0.79, 1.09]$ crosses the 1 cm bound. Restoring real-object context, unlike padding, is not consistently helpful (-3.0, $+1.7$, $+3.5$ cm). These are same-architecture replications on official room-level tests with known source overlaps excluded, not independent training seeds.

Continuous semantic channels. DiffuScene diffuses geometry, a ± 1 class one-hot block, and a 32-dimensional shape latent jointly under Gaussian noise; it has no mask state. On the same scenes, targets, lifts, and controls we map the conditions onto the class block of the noised state (Appendix A). Table 9 (Appendix B) shows that the ordering predicted by Equation 3 holds: at $t = 300$ the original-slot error rises from a certain correct class (18.4 and 16.5 cm) through the native noisy class (22.0, 19.6) and wrong classes ($+11.6$ and $+9.0$ cm over correct) to no information ($+16.4$ $[14.3, 18.6]$ and $+10.6$ $[9.5, 11.7]$) and the empty class ($+20.8$ and $+16.6$ cm). Pairwise ranking remains 1.000 at $t = 300$ and above 0.995 at $t = 150$, where the semantic contrasts are cleaner relative to the geometric noise floor (Appendix B).

Two MiDiffusion phenomena do not transfer, and both are properties of absorbing-state masking rather than of the mechanism. Wrong floor categories no longer retain most of the recovery; they cost 6–12 cm, within 2–6 cm of no information, because the asserted class conflicts with the target's native shape latent, which MiDiffusion does not have. And restoring the padding or real-object context changes the $t = 300$ estimate by at most 0.9 cm (-0.9 to $+0.8$ cm; at most 1.9 cm at $t = 150$), because Gaussian noise degrades but does not erase a padded slot's empty one-hot, so there is little occupancy information to restore; the discrete mask removes it entirely.

8 Related work

Scene generation and physical constraints. Layout generators range from autoregressive object sequences and sets (Wang et al., 2021; Paschalidou et al., 2021) through scene-graph and instruction conditioning (Zhai et al., 2023; Lin & Mu, 2024) to object-set diffusion and joint discrete–continuous diffusion (Tang et al., 2024; Hu et al., 2026). Physical outcomes are improved with collision and layout guidance (Yang et al., 2024), differentiable rigid-body simulation (Lin et al., 2026), or evaluator feedback at training and inference time (Wu et al., 2026); dynamics are grounded by explicit simulation (Xie et al., 2024; Liu et al., 2024; Zhang et al., 2024; Hu et al., 2019; Holl et al., 2020). PhyScene and PhyMix act on intermediate estimates during sampling, whereas PAT3D optimises a generated layout in a differentiable simulator; our results identify a condition the former must respect

in joint models, that such an estimate is an occupancy-weighted mean (§6). Our box-based floor test does not certify stability or interaction under forces.

Guidance, editing, and intervention scope. Classifier and classifier-free guidance steer sampling with classifier gradients or conditional–unconditional combinations (Dhariwal & Nichol, 2021; Ho & Salimans, 2022); SDEdit, ILVR, and diffusion posterior sampling impose reference or measurement information along the reverse process (Meng et al., 2022; Choi et al., 2021; Chung et al., 2023); self-guidance and stochastic self-guidance extract control signals from internal representations (Epstein et al., 2023; Chen et al., 2026). Latent editing, attention manipulation, and inversion connect representations to controllable changes (Kwon et al., 2023; Park et al., 2023; Hertz et al., 2023; Mokady et al., 2023). Our edits are diagnostics, not guidance methods; the categorical transition process (Austin et al., 2021) is why token edits persist while one-step continuous edits vanish.

Probing and denoising dynamics. Probing work distinguishes recoverable information from evidence about a model's computation (Belinkov, 2022); we intervene on inputs, not hidden units, and do not identify an internal mechanism. Analyses of diffusion dynamics characterise class commitment, and critical windows (Biroli et al., 2024; Sclocchi et al., 2025; Li & Chen, 2024).

9 Discussion and conclusion

What the results mean. A scene diffusion model that mixes discrete occupancy with continuous geometry cannot report an intermediate floor gap without also reporting its uncertainty about whether the slot exists. The affine identity makes this exact for posterior-mean estimators, the two heads agree quantitatively without a fitted parameter, and the step-resolved replays show the joint sampler dissolving the apparent violation as occupancy becomes readable, after which the token commits once. Nothing in this chain requires the model to have learned or failed to learn floor support; the finished samples support every retained floor object.

Implications. First, a physical diagnostic computed on an intermediate estimate measures occupancy unless it is conditioned on occupancy: absolute readouts should be reported together with the occupancy posterior, violations should be attributed to physics only for slots whose occupancy is resolved, and paired response scores are no evidence of a correct constraint plane. Second, guidance evaluated on the clean estimate must be gated on the model's own occupancy estimate: in every form whose edits persisted, a floor projection in the posterior mean or a separation or gradient step on the noisy state, edits that reached unresolved slots created objects. Gating removed most of these objects but did not reliably improve the targeted geometry, so guidance methods should be judged by their effect on occupancy as well as on the constraint they target. Third, the final support failures of the replayed checkpoints are occupancy and identity failures, never geometric ones, which points evaluation at the decomposition of final success into retention, category, applicability, and geometry, and remedies at the padding representation rather than at added support constraints.

Limitations. We study one constraint that is affine in the representation, one dataset, and two architectures. The cross-architecture evidence supports the mechanism, not the discrete token effects, and the DiffuScene interventions involve mild departures from the training input distribution. The guidance experiments are diagnostics with step rules fixed in advance, not comparisons with published guidance methods, and their collision and boundary measures are box proxies. Slot and category retention are not mesh identity, and native-generation references are retrospective. Pairwise constraints such as collision are not affine in a single slot's coordinates, and whether their intermediate readouts admit an analogous occupancy decomposition is open. We expect, but do not test, the same readout artefact in other mixed discrete–continuous generators with padded element sets and affine coordinate constraints, such as document and sketch layouts or floorplans.

Conclusion. Intermediate support violations in joint scene diffusion are occupancy uncertainty, not missing physics. Reading them as physical failures misdiagnoses the model, and correcting them without conditioning on occupancy creates the objects the model had not decided to place.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting sections of this paper and to improve the clarity, organization, and readability of the writing. We did not use generative AI tools to run the experiments. We reviewed and revised all AI-assisted text and checked technical claims, citations, and descriptions of experiments and results against the underlying work and original sources. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 2 states the affine-mixture identity and the parameter-free predictor, and Section 3 describes the checkpoints, the confirmation cohorts, the semantic interventions, the readouts and the uncertainty estimates. Appendix A gives the representations and target selection, the room splits and source coverage, the intervention and clipping rules, the inference settings and comparison units, the trajectory design and scene metrics, the third checkpoint, the step-resolved replay and guidance procedures, the DiffuScene condition mapping and the selection of the qualitative examples; Appendix B reports every condition and cohort behind the main-text tables; Appendix C derives the paired-score identity, the local response coordinates, the joint events and the posterior weight of an edited clean estimate; and Appendix D describes how the frozen protocols, source hashes and per-case outputs are organised. All experiments use the public 3D-FRONT dataset and the released MiDiffusion and DiffuScene checkpoints under their access conditions, with the seeds and shared random draws stated in Appendix A, and effects are reported with source-clustered bootstrap intervals.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 2022.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 2024.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *International Conference on Learning Representations*, 2026.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *International Conference on Computer Vision*, 2021.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Advances in Neural Information Processing Systems*, 2023.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3d furnished rooms with layouts and semantics. In *International Conference on Computer Vision*, 2021.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Philipp Holl, Vladlen Koltun, and Nils Thuerey. Learning to control PDEs with differentiable physics. In *International Conference on Learning Representations*, 2020.

Siyi Hu, Diego Martin Arroyo, Stephanie Debats, Fabian Manhardt, Luca Carlone, and Federico Tombari. Mixed diffusion for 3d indoor scene synthesis. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Yuanming Hu, Jiancheng Liu, Andrew Spielberg, Joshua B. Tenenbaum, William T. Freeman, Jiajun Wu, Daniela Rus, and Wojciech Matusik. ChainQueen: A real-time differentiable physical simulator for soft robotics. In *IEEE International Conference on Robotics and Automation*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations*, 2023.

Marvin Li and Sitan Chen. Critical windows: Non-asymptotic theory for feature emergence in diffusion models. In *International Conference on Machine Learning*, 2024.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. In *International Conference on Learning Representations*, 2024.

Guying Lin, Kemeng Huang, Michael Liu, Ruihan Gao, Hanke Chen, Lyuhao Chen, Beijia Lu, Taku Komura, Yuan Liu, Jun-Yan Zhu, and Minchen Li. PAT3D: Physics-augmented text-to-3d scene generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iIRxFkeCuY.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of Riemannian geometry. In *Advances in Neural Information Processing Systems*, 2023.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. DiffuScene: Denoising diffusion models for generative indoor scene synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. SceneFormer: Indoor scene generation with transformers. In *International Conference on 3D Vision*, 2021.

Dongli Wu, Jingyu Hu, Ka-Hei Hui, Xiaobao Wei, Chengwen Luo, Jianqiang Li, and Zhengzhe Liu. PhyMix: Towards physically consistent single-image 3d indoor scene generation with implicit–explicit optimization. *arXiv preprint arXiv:2604.10125*, 2026.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3D gaussians for generative dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically interactable 3d scene synthesis for embodied ai. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. CommonScenes: Generating commonsense 3d indoor scenes with scene graph diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3D objects via video generation. In *European Conference on Computer Vision*, 2024.

A EXPERIMENTAL DETAILS

A.1 REPRESENTATIONS AND APPLICABLE TARGETS

Continuous geometry encodes translation, half-extents, and yaw through the released model's normalization and angular representation. The categorical vocabulary contains real classes, an empty class, and a corruption mask. Empty and mask are distinct: empty denotes an unoccupied slot, whereas mask is a noisy semantic state. We use released room-specific vocabularies rather than sharing numeric category IDs across rooms.

The operational non-floor set consists of ceiling lamp, pendant lamp, wall lamp, television, and shelf when present in a vocabulary. A retained final category outside that set is floor-applicable under this evaluation. This rule does not certify every instance's true physical contact relation. Original target selection additionally requires positive geometry, an absolute floor gap at most 1 cm, and target geometry within encoding bounds. No output-dependent filtering selects targets or counterfactual categories. We do not exclude the main cohort for unrelated scene collisions or non-target clipping.

A.2 ROOM SPLITS AND SOURCE COVERAGE

Confirmation uses official test rooms with available processed assets and floor conditioning, excluding known source-UUID overlap with training and discovery cases. Bedroom has 107 scenes grouped into 103 source UUIDs; living room has 189 scenes and 189 UUIDs. The source mapping covers 3,879 of 6,286 official training bedroom rooms and 621 of 2,339 living rooms. This incomplete mapping prevents a house-independence claim even after removing known overlap. The older discovery cohort contains 43 official train, three validation, and four test rooms. Since the released configuration trains on train plus validation, 46 of the 50 discovery rooms are in its stated training splits. This conclusion uses the released configuration, not a recovered training log.

For confirmation, target selection covers different eligible categories where possible and then uses deterministic hash ordering, with at most three targets per scene. The resulting 309 and 567 targets yield $309 \times 16 \times 3 = 14{,}832$ and $567 \times 16 \times 3 = 27{,}216$ matched pairs at each primary noise level. Each pair contains both a lift and a support-preserving control; condition counts must not be confused with independent observations.

An archived 50-bedroom cohort, 46 of whose rooms lie in the released training splits, serves only for discovery and for the analyses in the appendices. Several design decisions were taken after earlier results had been seen and are recorded in the timing field of each frozen protocol in the supplementary material: the stratified analysis of §5 was motivated by an informal inspection of saved snapshots and then rerun under review, the every-step projection of §6 was adopted after a one-time projection proved inert, the footprint separation was designed after the floor-projection results, and the state-level variants after the separation inside the posterior mean had no detectable effect; each of these designs was frozen before its own outputs existed, except that the stratified protocol was frozen after the informal inspection had already shown the pattern it tests.

A.3 PRIMARY INTERVENTIONS AND CLIPPING

The target condition set is correct category, near-1, near-2, far-1, far-2, mask, and empty. Crossing these with clean and native context gives 14 conditions; adding untouched native semantics and four context-specific conditions gives 19. The four context conditions keep the target masked: restore other real-object tokens, restore padding tokens, or start from clean context and mask one real-object or padding token. One-slot selections are fixed before inference. Conditions unavailable because a scene has no padding are recorded as unavailable, not assigned an invented edit.

Incorrect categories are drawn from the room's floor-applicable vocabulary using official train-plus-validation statistics. Categories are ordered by standardized distances between median log-half-extents; two relatively near and two relatively far alternatives are selected without duplicates or the correct category. All four alternatives remain in the analysis.

For a positive lift δ, the target's center changes by δ. Its matched tangent control changes center and half-height by $\delta/\sqrt{2}$ each. These vectors have equal Euclidean norm in the physical (y, h) plane. Encoding may clip an input; stored actual gaps and displacements are consequently used in response

calculations. The unclipped-input stratum requires at most 10^{-6} m target encoding change in the lifted member used for MAE, not simultaneous unclipped status of both members. Control clipping is still recorded. This stratum coincides with the primary denominator in the present confirmation; the historical strict-valid subset has only 26 bedroom and 39 living-room scenes.

A.4 Inference, comparison units, and uncertainty

Both tested mixed models use 1,000 diffusion steps. We examine the native semantic process and crossed semantic interventions; the latter need not be typical samples from the native joint corruption distribution. Every comparison fixes the appropriate geometry, Gaussian realization, categorical draw, floor condition, and non-intervened tokens. Batch-size checks compare batches of 128 and 16 with a prespecified 2×10^{-5} tolerance in decoded gap (metres) or the corresponding model-coordinate outputs. They do not claim bitwise equivalence. Explicit identical-input and no-op gates check unintended edits.

For MAE, tail rates, and intervention effects, we average repeated targets, draws, and amplitudes within each scene; scenes then receive equal weight. PairAcc instead pools strict successes over pairs, so its non-unit values are pair-weighted. Bootstrap sampling draws source UUID clusters with replacement, retains their scenes, and recomputes the scene-equal effect for 2,000 resamples. Intervals are pointwise and are not adjusted for a large family of category comparisons. The prespecified four-category aggregate averages absolute errors and tail indicators per case before comparison to the correct category; it does not first average four geometric predictions.

All-sample and initially masked estimates are distinct. At $t = 300$ the invalid-member native target is masked in 4,584 of 14,832 bedroom pairs and 8,448 of 27,216 living-room pairs. Living-room one-padding interventions are available in 187 scenes and 26,928 pairs, so one-real versus one-padding comparisons use their shared denominator. Missing conditions are not counted as successful or failed predictions.

A.5 Trajectory design and scene metrics

Known-scene trajectories select one target per confirmation scene, apply a 10 cm lift, and use two independent draw combinations. Native generation instead uses two seeds on each of 32 deterministically selected floorplans per room. At $t = 300$, each of the four branches receives the same geometry and shares subsequent sampling randomness. Only target or context semantics are changed once. For native generation, the target is chosen by hash from baseline-final retained floor-applicable objects, not by selecting the largest baseline gap. Reference categories nevertheless come from that same final baseline and remain retrospective oracles.

All 301 reverse steps, including Python timestep zero, are checked for native/no-op equivalence; native-generation baseline continuations are also verified. All branch outputs are saved at timesteps 300, 250, 200, 150, 100, 50, and 0, together with their input states and predicted clean geometry. These seven snapshots do not constitute a complete saved semantic history.

Final occupancy follows the released decoder: the empty-class probability must be smaller than the maximum real-class probability for a slot to be retained. Final floor support requires that retained class to be applicable, all raw half-extents positive, and the absolute floor gap at most 5 cm. Collision is a rotated-box overlap proxy. Out-of-bounds evaluation uses the object footprint and the union of actual floor triangles. We also retain non-target support and object counts. None of these measurements is a mesh-level equilibrium or dynamic-stability test.

A.6 Dining-room cohort and third checkpoint

The dining-room cohort was built with the same prediction-blind rule: 175 official test scenes, 524 targets, 175 source UUIDs, 24 legacy strict-valid scenes, and 723 of 2,071 official training rooms mapped to a source. The execution and statistical protocol is identical to the bedroom and living-room wave; the runner's archived bedroom smoke was repeated after generalising it to a third room.

A.7 STEP-RESOLVED REPLAY AND PROJECTION GUIDANCE

Both experiments start from the saved $t = 300$ geometry, tokens, floor conditioning, and generator seeds or states of the trajectory cohorts and use a step-by-step replica of the adapter's joint sampler (network call, categorical posterior, Gumbel draw, then Gaussian draw). Every replayed branch that exists in the original runs reproduces the saved final slots and the seven saved snapshots with zero deviation. The replay records, for all 301 steps and all slots, the input and output tokens, the decoded clean-estimate gap, and the softmax of the auxiliary clean-class output. Projection edits the clean estimate's vertical centre to equal its half-height for selected slots, forms the mean of x_{t-1} from the edited estimate with $q(x_{t-1} \mid x_t, \hat{x}_0)$, and keeps the native variance and Gaussian draw; the gated arm selects slots whose auxiliary argmax over real classes plus empty is a floor-applicable real class. In the every-step design the edit at $t = 0$ fixes the selected slots' geometry by construction, so support and gap metrics of edited slots are reported but not interpreted as guidance quality. The footprint-separation replay uses the same machinery with a different edit: at every step with $300 \geq t \geq 50$, overlapping pairs among the selected slots (axis-aligned bounds of the yaw-rotated footprints overlap in x and z, the height intervals overlap, and the overlap volume exceeds $10^{-6}\,\mathrm{m}^3$) are moved apart along the horizontal axis of smaller penetration, each box by half the penetration plus 5 mm, with bounds recomputed after every move and at most five passes per step; only the x and z coordinates of moved slots are re-encoded, the gated arm selects slots whose auxiliary argmax is a real class, and the last 50 steps run natively so that the final geometry is not fixed by construction. The state-level variants reuse this separation. In the *state-consistent* form the displacement D of the clean estimate is also applied to the noisy state, holding the network's noise estimate fixed, so that the posterior mean becomes $q(x_{t-1} \mid x_t + \sqrt{\bar{\alpha}_t} D, \hat{x}_0 + D)$, which equals the native mean plus $\sqrt{\bar{\alpha}_{t-1}}\, D$ for the noise parameterisation of the released checkpoints; the edit confined to the posterior mean enters with weight $\beta_t \sqrt{\bar{\alpha}_{t-1}}/(1 - \bar{\alpha}_t)$ instead (0.006 against 0.63 at $t = 300$). In the *gradient* form, in the style of diffusion posterior sampling (Chung et al., 2023), the loss $\frac{1}{2} \sum \min(o_x, o_z)^2$ over overlapping selected pairs is differentiated with respect to the horizontal coordinates of the clean estimate, the gradient is propagated through the denoiser to the noisy state, and x_{t-1} is moved against it on the selected slots with the per-axis step $\frac{1}{2} \sqrt{\bar{\alpha}_t \bar{\alpha}_{t-1}}/h_c^2$, where h_c is the metric half-range of axis c; if the network's noise prediction were held fixed, this step would move each box of an overlapping pair by half its penetration, so differences between the two forms reflect the denoiser's response to the edit, and also the loss (a translation gradient vanishes for nested or coincident footprints, and overlapping pairs are summed in one step rather than separated in passes with a 5 mm margin). Neither step size was tuned: both follow from these derivations and were fixed before any output existed. The gate is recomputed at every step, and the last 50 steps run natively in all four arms.

A.8 DIFFUSCENE CONDITION MAPPING

DiffuScene encodes each slot as translation, half-extents, yaw, a ± 1 class one-hot with a final empty class, and a 32-dimensional shape latent, and pads absent slots with zero geometry and the empty one-hot. One Gaussian draw over all channels is shared by both pair members, all lifts, all conditions, and both timesteps. A *certain* class sets the target's class block to that one-hot and removes its class-block noise; *no information* sets the class block and the shape-latent signal to zero while keeping the native noise; clean context removes class-block noise on every other slot; restore-real and restore-padding do so on real or padding slots only; mask-one conditions start from the clean context and remove the information of one hash-selected slot with the same key as the MiDiffusion conditions. Wrong-class conditions therefore keep the target's native shape latent, which conflicts with the asserted class; this is a stated confound of the wrong-class contrast, not of the no-information or empty contrasts. *Native target masked* for this architecture means that the argmax of the noisy native class block differs from the original class (about 60% of pairs at $t = 300$, 4.5% at $t = 150$). Batch equivalence within 2×10^{-5} required disabling TF32 for cuDNN and cuBLAS and enabling deterministic convolution algorithms; with default settings the estimates differed by 3×10^{-3} between batch sizes.

A.9 QUALITATIVE EXAMPLES

The two examples of Figure 1 were selected by the smallest stable hash of the case or trajectory identifier within a prespecified eligible set, not by appearance. For the single-step panels the eligible set is the natively masked bedroom targets with a 10 cm lift, Gaussian draw 0, and categorical draw

Table 4: Full primary matrix at $t = 300$: scene-equal original-slot absolute gap (cm). All 19 conditions have exact PairAcc 1 in each checkpoint. Empty rows are representation readouts, not errors of occupied floor-supported objects. All rows use 107/189/175 scenes except living-room mask-one-padding, which uses 187.

Target	Context	Bedroom	Living room	Dining room
Native	Native	17.347	13.838	13.478
Correct	Clean	2.323	2.601	2.347
Wrong near-1	Clean	2.859	3.051	3.132
Wrong near-2	Clean	3.171	3.126	2.927
Wrong far-1	Clean	2.570	3.264	3.544
Wrong far-2	Clean	3.671	3.496	3.393
Mask	Clean	27.381	20.862	25.418
Empty	Clean	93.104	80.580	81.396
Correct	Native	2.247	2.397	2.239
Wrong near-1	Native	2.774	2.952	3.030
Wrong near-2	Native	3.026	2.897	2.691
Wrong far-1	Native	2.338	3.179	3.532
Wrong far-2	Native	3.351	3.482	3.474
Mask	Native	50.640	38.972	38.886
Empty	Native	93.515	80.484	81.429
Mask	Restore real context	47.665	40.657	42.433
Mask	Restore padding context	37.834	27.152	27.598
Mask	Mask one real (from clean)	36.657	24.691	26.279
Mask	Mask one padding (from clean)	40.330	29.141	32.772

100 (96 cases); for the final-sample panels it is the bedroom native-generation trajectories in which the occupancy-blind arm sampled more objects than the native branch (34 of 64). The selected target is a nightstand whose clean estimate floats 75.2 cm above the floor with an empty posterior of 0.84 under the masked token and lies within 0.1 cm of the floor with an empty posterior below 0.001 once the token is restored. The selected trajectory has five objects and one overlapping box pair under native continuation and six objects with six overlapping pairs under occupancy-blind projection; the added object is a pendant-lamp token with a large footprint, and the gated arm reproduces the native counts. Only slots kept by the decoder are drawn in the final-sample panels.

B Additional results and denominators

B.1 Complete primary semantic matrix

Table 4 retains every primary condition. The strong effect of target restoration does not imply that exact category identity is irrelevant. For far-2, the MAE excess over correct is 1.104 cm with interval $[0.555, 2.096]$ in bedrooms and 1.085 cm $[0.920, 1.253]$ in living rooms. In the bedroom originally masked subset, the four-category aggregate excess is 0.856 cm $[0.498, 1.400]$, also outside the prespecified interval-containment criterion. The corresponding living-room excess is 0.721 cm $[0.593, 0.848]$.

Restoring all context with the target fixed to mask changes MAE by -23.259 cm $[-26.267, -20.059]$ and -18.110 cm $[-21.160, -14.707]$ relative to native context. In the clean-context one-slot control, one-real minus one-padding gives -3.673 cm $[-6.657, -0.719]$ and -4.983 cm $[-6.893, -2.988]$, respectively. The living-room contrast uses 187 mutually eligible scenes; subtracting its two full-table means would use different denominators. These results establish context dependence without measuring per-token importance or assuming additive effects.

B.2 Lamp conditions and semantic uncertainty

With clean context, ceiling- and pendant-lamp original-slot coordinates are 226.389 and 169.069 cm in bedrooms, and 226.425 and 176.685 cm in living rooms. With native context, the ceiling-minus-

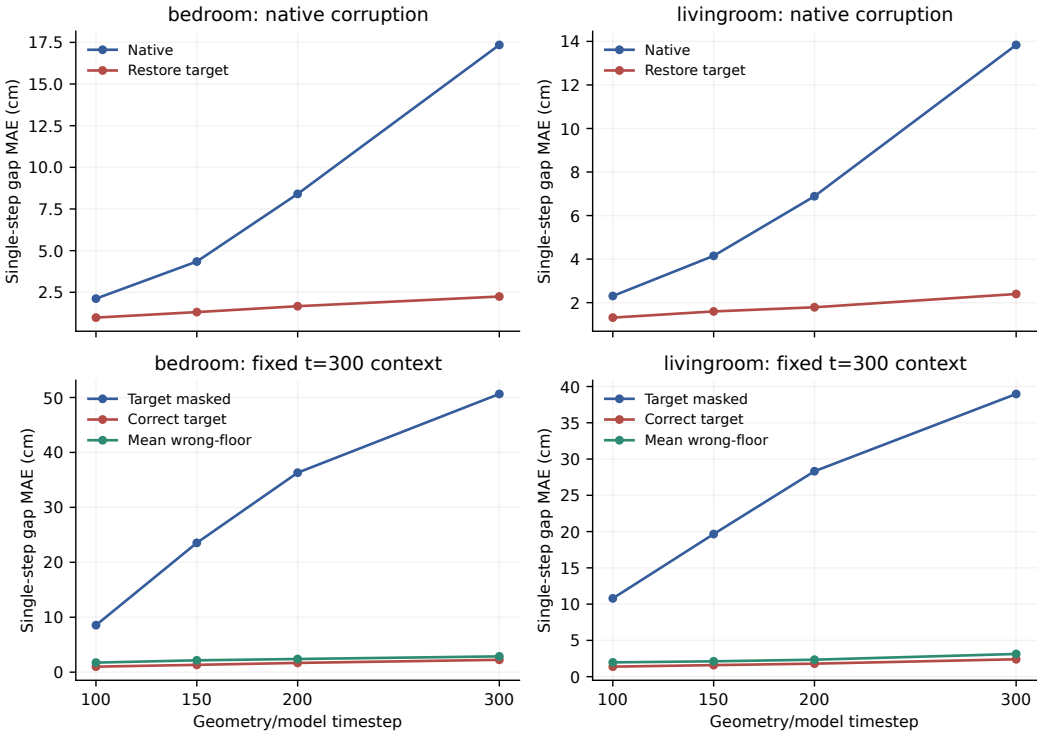


Figure 3: Scene-equal single-step original-slot readouts on the frozen confirmation cohorts. Fixed semantic conditions reveal a residual noise/timestep dependence; native mask frequency is not the sole source of the trend. Lines are descriptive means; matched effect intervals are reported separately in the text.

correct contrasts are 224.283 cm $[221.182, 227.211]$ and 226.142 cm $[223.627, 228.504]$; pendant-minus-correct gives 167.718 cm $[164.194, 171.067]$ and 176.573 cm $[173.693, 179.357]$. These are coordinate changes under a changed semantic constraint, not lamp-failure rates.

In the earlier 50-scene native $t = 300$ readout, 143 cases have actual original-slot gap above 20 cm. Of these, 85 have auxiliary semantic argmax empty, four ceiling/pendant lamp, and 54 other furniture categories. These auxiliary predictions are not final sampled classes and are not established mediators of the geometric effect.

B.3 Noise curves

Figure 3 compares native, restored-target, and fixed-mask conditions. The fixed-context curves use the same $t = 300$ categorical realization at other geometric timesteps and hence are crossed input interventions. In bedrooms, native MAE at $t = 100, 150, 200, 300$ is 2.120, 4.345, 8.406, and 17.347 cm; correct-target restoration gives 0.981, 1.311, 1.664, and 2.247 cm. The corresponding living-room sequences are 2.305, 4.150, 6.887, 13.838 and 1.313, 1.597, 1.787, 2.397 cm. These curves do not isolate geometric corruption from the network's timestep input.

B.4 Matched final gaps and scene-level effects

Table 5 uses only targets retained and floor-applicable in both the target-restored and native branches. The living-room forward subset includes 187 scenes, not all 189. Its conditional comparison must remain separate from the 378-trajectory joint-success denominator.

For known scenes, target restoration changes the fraction of colliding objects by $+1.463$ percentage points $[-0.191, 3.308]$ in bedrooms and $+0.522$ $[-0.350, 1.350]$ in living rooms. Boundary violations change by $+0.923$ $[-0.973, 2.739]$ and $+0.069$ $[-0.395, 0.550]$. Context restoration changes

Table 5: Final gap MAE on the same surviving floor-applicable targets in both branches. Values and differences are cm. n counts matched trajectories; intervals are the original pointwise source-cluster bootstrap intervals. This is an outcome-defined subset.

Room	Source	n	Native	Target once	Difference [95% CI]
Bedroom	Known scene	195	0.17722	0.17467	-0.00255 $[-0.01026, 0.00605]$
Living room	Known scene	343	0.17669	0.18379	$+0.00709$ $[-0.00076, 0.01555]$
Bedroom	Generation	64	0.13227	0.13538	$+0.00311$ $[-0.01446, 0.02406]$
Living room	Generation	64	0.21838	0.21316	-0.00521 $[-0.03722, 0.03274]$

the bedroom collision fraction by -6.115 $[-11.151, -1.215]$; this reduction is not replicated in living rooms. Thus we do not summarize every context effect as harmful or every target effect as beneficial to the whole scene.

For native generation, context restoration gives retention/floor-applicability/supported-occupancy rates of 96.875% in bedrooms and 98.438% in living rooms, while category-correct and joint rates are both 89.063%. In this sample the joint losses therefore do not arise from retained applicable objects crossing the 5 cm support threshold. The baseline ceiling and reference selection remain part of this interpretation.

B.5 Frozen mixture confirmation

The mixture uses only noisy target height, half-height, and semantic token. Class-conditioned two-dimensional Gaussians and slot-frequency priors are estimated from official training-plus-validation geometry without fitting neural predictions. The seven fitted variants are empirical all-class with zero padding; empirical real-class without empty; empirical floor-only without empty; constrained floor-only without empty; constrained floor with zero padding; constrained floor plus empirical non-floor classes and zero padding; and the latter with support-compatible padding. All variants are retained; none is selected by agreement with the network. Separate synthetic mixtures illustrate the affine identity and are not one of these seven fitted models.

On the bedroom test cohort at $t = 300$, the empirical real-class mixture without empty yields native MAE 10.486 cm and per-case network-gap RMSE 20.951 cm. Restricting to empirical floor classes gives 0.102 cm MAE but RMSE 33.871 cm. These examples show why reproducing one aggregate error is not sufficient. The strictly floor-supported, empty-free construction preserves the affine constraint up to numerical precision. No living-room mixture is fitted, and no baseline receives context tokens.

B.6 One-time projection

Applying the projection once at $t = 300$ leaves every final metric unchanged to within 10^{-5} m: in the bedroom native-generation cohort the projection acts on all 12 slots (mean maximum gap 1.62 m before, 10^{-7} m after), the next clean estimate changes by 0.015 cm, and the difference decays to 10^{-4} cm by $t = 0$, with a final raw deviation of 4×10^{-4}. The DDPM posterior mean weights the edited estimate by $\beta_t \sqrt{\bar{\alpha}_{t-1}}/(1 - \bar{\alpha}_t) = 0.0063$ at $t = 300$.

B.7 Footprint separation in the posterior mean and in the state

Table 6 gives the scene-equal paired differences of the every-step footprint separation (§6); every interval in the table includes zero. In the native-generation cohorts both arms lose 1.6–3.1 points of target retention or category agreement with intervals reaching zero, as in the floor-projection experiment. The first edit at $t = 300$ performs on average 88 and 193 pair separations per scene (pair moves summed over up to five passes) in the occupancy-blind arm of the known-scene cohorts (86

Table 6: Every-step footprint separation of overlapping boxes in the clean estimate for $300 \geq t \geq 50$ (last 50 steps native), guided minus native (scene-equal, 95% cluster-bootstrap intervals). Objects: sampled objects per scene; collisions, boundary: colliding and out-of-bounds object fractions in percentage points.

Cohort	Arm	Objects	Collisions (pp)	Boundary (pp)
Bedroom, known (107)	occupancy-blind	$+0.02$ [-0.01, 0.06]	-1.1 [-3.2, 1.1]	$+1.1$ [-0.3, 2.6]
	gated	$+0.00$ [-0.01, 0.01]	-0.4 [-1.6, 0.7]	$+0.3$ [-0.2, 0.9]
Living room, known (189)	occupancy-blind	$+0.00$ [-0.02, 0.02]	-0.3 [-0.9, 0.2]	$+0.1$ [-0.2, 0.4]
	gated	$+0.00$ [-0.01, 0.01]	-0.1 [-0.6, 0.3]	-0.1 [-0.3, 0.1]
Bedroom, gen. (32)	occupancy-blind	$+0.03$ [0.00, 0.08]	$+0.3$ [-2.2, 2.7]	$+0.9$ [-0.9, 3.1]
	gated	-0.02 [-0.05, 0.00]	-1.1 [-2.5, 0.3]	$+0.0$ [-2.4, 2.0]
Living room, gen. (32)	occupancy-blind	$+0.03$ [-0.03, 0.09]	-0.1 [-1.6, 1.5]	$+0.1$ [-0.9, 1.3]
	gated	-0.02 [-0.05, 0.00]	-0.4 [-1.4, 0.5]	$+0.2$ [-0.4, 0.9]

Table 7: Footprint separation carried into the noisy state (state: the displacement of Table 6 applied with the noise estimate held fixed; grad.: penetration-loss gradient through the denoiser), $300 \geq t \geq 50$ with the last 50 steps native, guided minus native (scene-equal, 95% cluster-bootstrap intervals). Objects: sampled objects per scene; collisions: colliding-object fraction; category: target category agreement, both in percentage points.

Cohort	Arm	Objects	Collisions (pp)	Category (pp)
Bedroom, known (107)	state, blind	$+1.06$ [0.92, 1.21]	$+9.1$ [4.5, 13.6]	-5.1 [-9.3, -1.4]
	state, gated	$+0.12$ [0.05, 0.19]	-2.8 [-6.9, 1.2]	-0.9 [-3.3, 1.4]
	grad., blind	$+0.35$ [0.27, 0.43]	$+0.7$ [-3.0, 4.9]	$+0.9$ [0.0, 2.4]
	grad., gated	$+0.01$ [-0.02, 0.05]	-2.4 [-5.0, 0.3]	$+0.9$ [0.0, 2.4]
Living room, known (189)	state, blind	$+0.36$ [0.24, 0.50]	$+7.5$ [5.2, 9.9]	-0.3 [-1.9, 1.6]
	state, gated	$+0.06$ [0.01, 0.10]	-1.2 [-2.6, 0.2]	$+0.5$ [-0.8, 2.1]
	grad., blind	$+0.23$ [0.18, 0.29]	-0.4 [-1.7, 0.9]	-0.5 [-2.1, 1.1]
	grad., gated	$+0.00$ [-0.02, 0.02]	$+0.4$ [-0.4, 1.2]	$+0.5$ [-0.5, 1.6]
Bedroom, gen. (32)	state, blind	$+1.09$ [0.84, 1.33]	$+6.3$ [-1.6, 14.2]	-25.0 [-37.5, -14.1]
	state, gated	$+0.11$ [-0.02, 0.23]	-3.4 [-8.6, 1.9]	-12.5 [-20.3, -6.2]
	grad., blind	$+0.33$ [0.16, 0.53]	$+0.4$ [-5.4, 6.5]	-9.4 [-17.2, -3.1]
	grad., gated	-0.02 [-0.08, 0.03]	-2.0 [-5.5, 1.0]	-3.1 [-7.8, 0.0]
Living room, gen. (32)	state, blind	$+0.06$ [-0.22, 0.36]	$+8.9$ [2.7, 15.6]	-9.4 [-17.2, -3.1]
	state, gated	$+0.05$ [0.00, 0.11]	$+2.8$ [-0.7, 6.1]	-10.9 [-18.8, -4.7]
	grad., blind	$+0.25$ [0.11, 0.39]	$+1.0$ [-1.2, 3.4]	-1.6 [-4.7, 0.0]
	grad., gated	$+0.02$ [0.00, 0.05]	$+1.2$ [-0.9, 3.0]	-1.6 [-4.7, 0.0]

and 192 in the generated cohorts), against 11 and 37 (6 and 34) in the gated arm, which separates only real-argmax slots.

Table 7 gives the two state-level forms. The prespecified criterion, a gated colliding-object interval entirely below zero together with an object-count interval that includes zero in both known-scene cohorts, is met by neither form: the state form's gated arm adds 0.12 $[0.05, 0.19]$ and 0.06 $[0.01, 0.10]$ objects in those cohorts, and no gated colliding-object interval lies below zero. Boundary violations change by -2.2 to $+3.6$ points, with one interval excluding zero (bedroom, known scenes, occupancy-blind gradient: $+3.4$ $[0.7, 6.2]$). In the native-generation cohorts the state form lowers target retention by 1.6–7.8 points in both arms. The protocol's colliding-pair count changes by -0.12 to $+0.34$ pairs per scene in the gated arms, with one interval excluding zero (bedroom, generated scenes, gated gradient: -0.11 $[-0.20, -0.03]$). The gradient form's per-case edit counts are not comparable with the state form's, because attention spreads the gradient from the overlapping pairs to other selected slots.

B.8 Regression of the empty posterior on masked context slots

Pooled and within-scene ordinary least squares of the masked target's empty posterior on the numbers of masked real-object and masked padding context slots, with source-cluster bootstrap intervals: bedroom within-scene $+0.060$ $[0.052, 0.066]$ per padding slot and -0.012 $[-0.024, -0.001]$ per

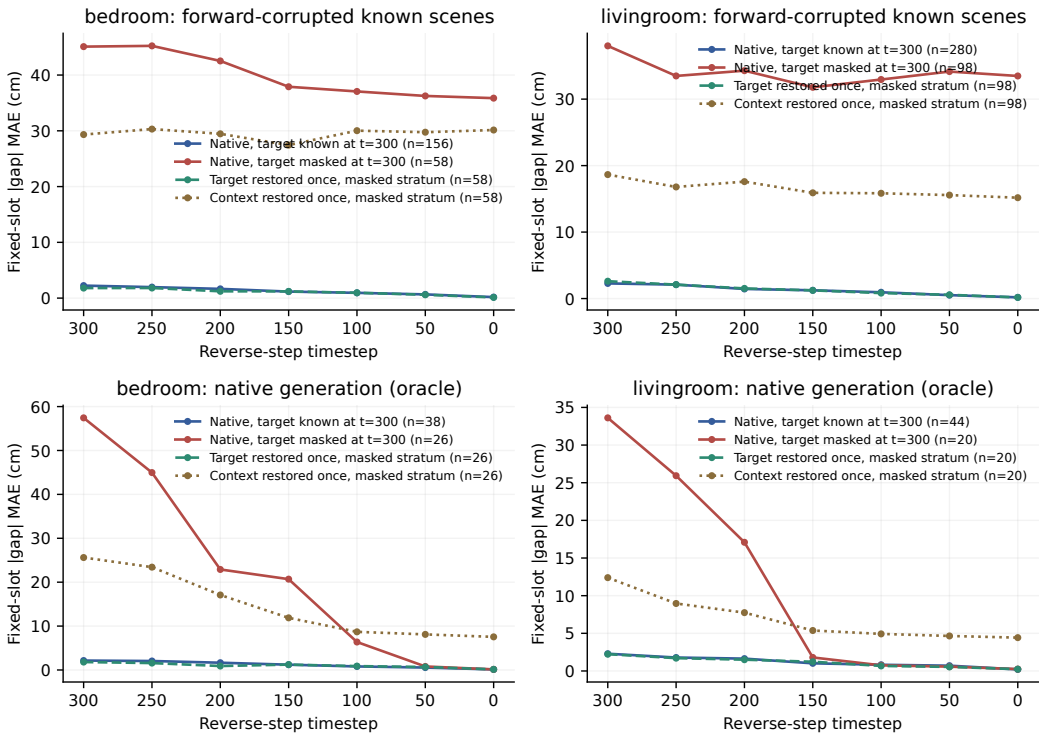


Figure 4: Scene-equal original-slot readouts along the saved snapshots, stratified by the native $t = 300$ token of the target. Curves include slots that later become empty or change class and are therefore representation readouts, not final-object errors.

Table 8: Step-resolved replay of masked targets under native continuation. "Geometry first": the clean estimate enters the 5 cm band before the input token leaves <mask>; steps count down from 300; medians are over trajectories with the event.

Cohort	n	Geom. first	Same	Token first	Never ≤ 5 cm	Unmask	Settle
Bedroom, known scenes	58	28	10	1	19	160	210
Living room, known scenes	98	55	11	1	31	127	195
Bedroom, generation	26	16	10	0	0	127	176
Living room, generation	20	13	6	1	0	174	185

real slot; living room $+0.048$ $[0.043, 0.052]$ and -0.019 $[-0.024, -0.014]$. The corresponding gap coefficients are $+4.9$ $[3.8, 5.9]$ and $+2.8$ $[2.3, 3.3]$ cm per masked padding slot. The natively masked subset of the unmodified condition gives the same signs and magnitudes.

B.9 Deleted targets are visible at the intervention step

Among masked targets under native continuation, those finally deleted have a mean auxiliary empty posterior of 0.556 at $t = 300$ in both rooms (median clean-estimate gap 65 and 47 cm), those retained with another class 0.31 and 0.27 (42 and 25 cm; 20 of 22 and 45 of 52 floor-supported), and those retained with the reference class 0.19 and 0.33 (18 and 31 cm). Deleted targets output <empty> for a median of 166 and 133 of the 301 steps and never settle within the 5 cm band.

B.10 Stratified intermediate readouts and step-resolved ordering

Medians of the unmask step are over all masked targets with the event; settle medians are over retained targets. In 720 native trajectories the output token of a masked target changes exactly once

Table 9: DiffuScene class-channel interventions on the frozen cohorts: scene-equal original-slot gap MAE (cm) and paired differences with 95% cluster-bootstrap intervals. "No information" removes the target's class and shape signal; certain classes are noise-free one-hots.

t	Contrast	Bedroom	95% CI	Living room	95% CI
300	native MAE / certain correct MAE	22.0 / 18.4		19.6 / 16.5	
300	certain correct − native	−3.60	[−5.43, −2.04]	−3.12	[−4.08, −2.15]
300	wrong-class average − correct	+11.63	[10.42, 12.90]	+8.97	[8.27, 9.77]
300	no information − correct	+16.40	[14.32, 18.62]	+10.62	[9.54, 11.73]
300	empty − correct	+20.77	[19.23, 22.25]	+16.55	[15.84, 17.37]
300	padding restored − no information	+0.35	[−0.03, 0.72]	+0.77	[0.47, 1.07]
150	native MAE / certain correct MAE	2.6 / 2.2		5.3 / 4.0	
150	wrong-class average − correct	+6.02	[4.82, 7.46]	+5.68	[4.62, 6.85]
150	no information − correct	+11.65	[9.15, 14.57]	+11.74	[9.87, 13.84]
150	empty − correct	+27.37	[24.63, 30.32]	+25.15	[23.25, 27.19]

Table 10: Diagnostic counterexamples on the 600-pair exploratory construction, in scene coordinates with floor $f = 0$. Tangent change is the RMS change in $(y + h)/\sqrt{2}$; all geometric values are in cm. The wrong-plane example at 20 cm retains positive target half-heights in these cases.

Mapping	Gap MAE	PairAcc	Tangent change
Projection onto the floor	0	1.000	0
Projection onto the plane $g = 20$ cm	20	1.000	0
Fixed vertical output $y = h = 0.5$ m	0	1.000	49.99
Identity	12.47	0.000	0

and no trajectory switches between two real classes; none of the 518 known-stratum targets changes token in 301 steps.

B.11 DiffuScene contrasts

C Diagnostic identities, coordinates, and sampler scope

C.1 Shared offsets cancel in the paired score

Replacing both predicted gaps by $\hat{g}_j + b$ leaves Equation 4 unchanged. In particular, $\hat{g}_I = \hat{g}_C = b$ yields $S_I - S_C = q$ for a positive lift with actual input difference q. This result does not require a denoiser to inspect either input. A wrong-plane projector and an input-independent constant are therefore complementary counterexamples: the first preserves support-tangent variation but locates the wrong plane; the second can return the correct gap while discarding that variation.

In the exploratory $t = 300$ native data, the lifted and control members have MAE 15.613 and 15.570 cm, while their output-gap difference has RMS 0.470 cm. Single-response fitting against the nominal injected gap gives $R^2 = 0.000546$; subtracting the matched control gives $R^2 = 0.998378$. The corresponding actual-input-gap fit gives $R^2 = 0.998709$. Thus strong differential fit can coexist with a large shared absolute offset. These fits use the historical no-intercept convention and are not an independent confirmation experiment.

C.2 Local response coordinates

Write $z = (y, h)$ for the varied physical coordinates, E for their affine model encoding, and D for decoding. Other geometry and semantics are fixed. Away from clipping, the measured composite map is

$$F_t(z; c, \epsilon) = D\big(\widehat{x}_0\big(\sqrt{\bar{\alpha}_t}\, E(z) + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, c, t\big)\big)\,. \tag{5}$$

Here $\widehat{x}_0$ denotes the relevant geometric outputs, with the unvaried inputs implicit. If A is the encoding's linear part, its derivative contains $\sqrt{\bar{\alpha}_t} A^{-1} J_{\widehat{x}_0} A$, not simply the denoiser Jacobian. This

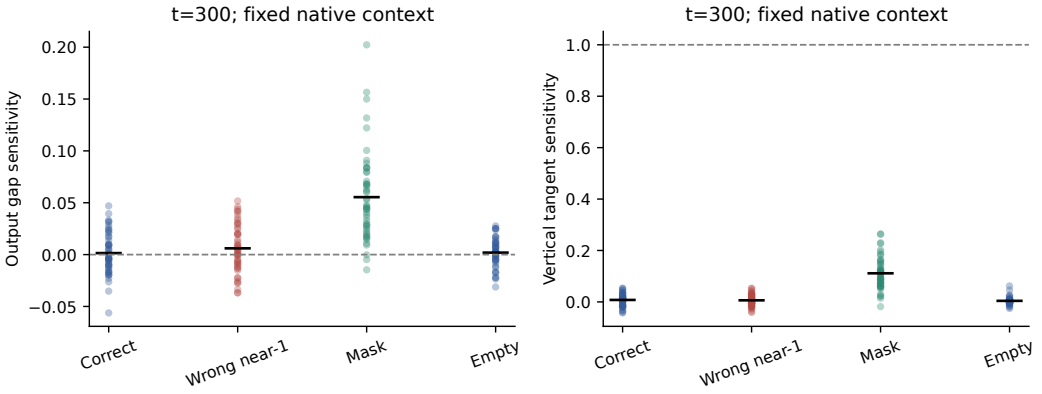


Figure 5: Finite-perturbation slopes at $t = 300$ on the exploratory 50-scene cohort, fixed native $t = 300$ context, and unclipped positive-size inputs. The plotted "normal" label refers to a vertical lift and output-gap slope, not a pure-normal Jacobian eigenvalue. Tangent contraction limits an interpretation as support projection.

restriction excludes the yaw-to-sine/cosine representation, which is not a square affine transformation. Physical and normalized Euclidean metrics also differ.

The artifact's "normal" curve varies y only. This direction has both normal and tangent components in the (y, h) plane; its output-gap slope is not a normal eigenvalue. The tangent curve changes both coordinates by $\delta/\sqrt{2}$, preserves g, and measures output $(y + h)/\sqrt{2}$ against its actual input value. An ideal projection onto $g = 0$ would give gap slope zero and tangent slope one under this test. Their converse does not identify a full projection, and our finite fitted slopes do not estimate a zero-amplitude Jacobian.

The signed grid is $0, \pm5, \pm10, \pm15, \pm20$ cm. Fits are retained both for all encoded inputs and for unclipped positive-size inputs. In the latter stratum, clean-context correct-target fits at $t = 1, 10, 50$ have gap slopes 0.20748, 0.00792, and -0.00058, tangent slopes 0.48665, 0.11552, and 0.01747, and zero-perturbation gap MAE 0.148, 0.241, and 0.628 cm. The $t = 1$ gap-fit residual RMS is 1.302 cm, underscoring finite-range nonlinearity. These low-noise timesteps were added after examining the $t = 150/300$ curves and are exploratory.

C.3 Affine mixtures and empty representations

For $g(z) = a^\top z + b$ and any integrable conditional distribution, linearity gives $g(\mathbb{E}[z \mid u]) = \mathbb{E}[g(z) \mid u]$. If all supported states lie in $g = 0$, their mixture mean does too, irrespective of category probabilities or sizes. A zero vector in normalized geometry need not decode to this plane. For the bedroom bounds,

$$g(D(0)) = \frac{\ell_y + u_y}{2} - \frac{\ell_h + u_h}{2} - f \simeq 93.99\,\text{cm}. \tag{6}$$

This is a coordinate consequence of padding, not a force or a physical object. Adding this representation or non-floor states to a toy prior can shift its conditional mean. Re-encoding a toy empty state to satisfy $g = 0$ restores the affine identity; this intervention changes the toy prior, not the trained neural checkpoint.

C.4 Joint events and the semantic sampler

Because reference target categories are floor-applicable, C implies A. Consequently $J = C - C(1 - H)$ case by case. In the saved trajectory sample, $C(1 - H) = 0$ and $A(1 - H) = 0$ for every branch. This proves the empirical event identities used in the main text under the chosen threshold, without establishing a mediation decomposition. Separately scene-weighted conditional means cannot in general be multiplied to recover scene-weighted joint rates.

The mixed sampler uses the predicted clean-class distribution in its categorical posterior for positive timesteps. At Python timestep zero it instead samples that predicted distribution directly. The configured cumulative keep and mask coefficients sum to 0.999999, leaving approximately 10^{-6} total uniform replacement mass over non-mask classes, including empty and self-redraw. Finite log-probability floors are also used. Forward replacement mass is not a reverse category-switch probability, and numerical floors do not establish an observable switch rate. Most importantly, there is no hard target-copy rule in these one-time interventions; the upstream partial-class clamping path is not used.

For target-restored branches, all seven saved output tokens match the reference, with zero observed input-to-output switches at the six saved positive timesteps and at timestep zero. Final correct counts are 214/214 and 378/378 for known scenes and 64/64 per room for native generation. These observations cannot exclude switches and reversions between snapshots, prove persistence across all 301 steps, or determine what fraction of the benefit is due to the transition kernel.

C.5 Posterior weight of an edited clean estimate

For the linear schedule $\beta_t \in [10^{-4}, 0.02]$ with 1,000 steps, the coefficient of $\hat{x}_0$ in the DDPM posterior mean, $\beta_t\sqrt{\bar{\alpha}_{t-1}}/(1-\bar{\alpha}_t)$, equals 0.0063 at $t = 300$ and equals 1 at $t = 0$, where the coefficient of x_t vanishes. A one-time edit of the clean estimate at $t = 300$ therefore perturbs x_{299} by less than one percent of the edit, whereas an edit at the final step is copied into the sample; this is why the every-step design of §6 fixes the selected slots' geometry by construction at $t = 0$ and why the one-time design has no effect.

D Reproducibility and historical evidence boundaries

Historical response evidence. An earlier exploratory study measured support-related responses in DiffuScene and MiDiffusion across bedroom, living-room, and dining-room checkpoints. A mixed-model response replication covered 29,100 matched pairs across these room types and reported perfect ranking, with low-noise contraction coefficients λ near one in its no-intercept fit $\Delta g \approx -\lambda g$. These coefficients are not the output-gap slopes of Appendix C.2. They motivated the present evaluation but do not demonstrate tangent preservation or independently confirm semantic interventions. Historical random validation splits must also not be confused with the new official room-test cohorts.

The earlier work examined variance, branching, and collision-related readouts. Such results concern different measurements and do not supply missing evidence for semantic intervention mechanisms or final scene improvement. We do not reuse the archived passive-label classification table whose label provenance remains unresolved. The present conclusions rely on the explicit paired identities, new semantic confirmation, and saved joint trajectories.

Artifact structure. Each experiment records its cohort and seed manifest, exact command, environment, checkpoint and source hashes, input provenance, verification results, and completed or failed status. Primary semantic runs retain per-case predictions, auxiliary logits, original-slot geometry, and shared noise/token inputs. Trajectory runs retain final sampled slots for all branches, seven input/output snapshots, per-case final geometry metrics, and the full-step no-op verification. Analyses retain scene-level effects, pointwise intervals, and their denominators. Every table and figure is transcribed from these records.

Review and analysis status. After the review protocol was adopted, new experiment and analysis code was independently reviewed before execution, with approvals tied to source hashes and execution scope, and reviewed again before commits. The initial historical-output audit preceded that protocol and is not retroactively described as pre-reviewed. These agent-based reviews are execution and consistency checks, not independent model replications. The primary confirmation protocol was frozen before those predictions. Lamp-probe details and the low-noise local expansion were specified after earlier results, while the final event decomposition and semantic-snapshot audit are explicitly post-result analyses. The audit reused existing intervals without drawing a new bootstrap. Identical events can consequently have slightly different archived interval endpoints because the original bootstrap seed depended on the metric name.

Release preparation. The supplementary material contains the code of this study, of the discovery replay on the archived bedroom cohort, and of the earlier exploratory study, together with the frozen protocols; identifying paths are removed. Experiment outputs, datasets, and checkpoints are not included; the datasets and checkpoints are available from their original sources under their access conditions.