LEARNING PREDICTIVE SCENE REPRESENTATIONS FROM RAW 3D GAUSSIAN SPLAT PARAMETERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gaussian splatting has become a common output of scene capture, and representation learning on splats has mainly distilled image features or reconstructed Gaussian parameters. Inferring what a scene hides, or anticipating what an edit will change, requires a representation that is predictable from context. We learn such representations directly from raw Gaussian parameters by predicting the latent representations of masked scene regions from their visible surroundings; images supply training targets only, so inference needs neither photographs nor rendering. Measuring recognition and prediction separately reveals that they are distinct and that image supervision trades one for the other: stronger supervision improves recognition while steadily lowering predictability, and features lifted from photographs, although better recognizers, are far less predictable than ours under matched predictors. Adding relative spatial position to the predictor makes its predictions depend on how a scene is arranged. These representations enable two capabilities: predicted embeddings retrieve real Gaussian groups from other scenes to fill hidden regions, and an action-conditioned latent world model, trained separately on synthesized edits, predicts how a scene's representation changes when objects are removed, moved, or added. Its removal predictions improve steadily with more training scenes, and searching over its predictions recovers the edits that produced a goal scene far more often than chance. Captured splats can thus be used to predict, not only to render.

1 INTRODUCTION

3D Gaussian Splatting [3] has become a common output of scene capture. Existing methods learn splat representations through image-feature distillation [4] or Gaussian-parameter reconstruction [6]. We keep Gaussian parameters as input and study a different objective: predicting masked contextual embeddings. This lets us ask which properties a downstream system can infer from the splat alone.

A captured room provides a concrete prediction task: infer the representation of a hidden region, or anticipate how that representation changes when an object is edited. Both tasks need the surrounding scene, not only the object. Predictive learning has made representations useful for related tasks in video [41]; here we study captured splats. We evaluate both *recognition*, whether a linear probe recovers semantic classes after pretraining without class labels, and *predictability*, whether visible context supports inference of hidden representations. Recognition accuracy alone misses a predictive capability that we measure directly: in controlled sweeps, stronger image supervision raises recognition while lowering predictability.

We train a joint-embedding predictive architecture (JEPA) [10; 58; 59; 11] on groups of raw 59-dimensional Gaussian parameters (Figure 1). Given the visible groups of a scene, a predictor regresses the encoder's embeddings of the masked groups, and a distributional regularizer (SIGReg-style [11], or a VISReg-inspired scale–shape hybrid [13] for our main model) prevents collapse in place of a momentum teacher. In our main model, each group is represented by a geometry token and an appearance token; only the appearance token is additionally trained to match frozen image features lifted from the scene's photographs, so images are used to build training targets and never at inference. Matched null baselines, including copying a visible token, test what prediction adds beyond them. On the frozen representation we train an *action-conditioned latent world model*: it

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.

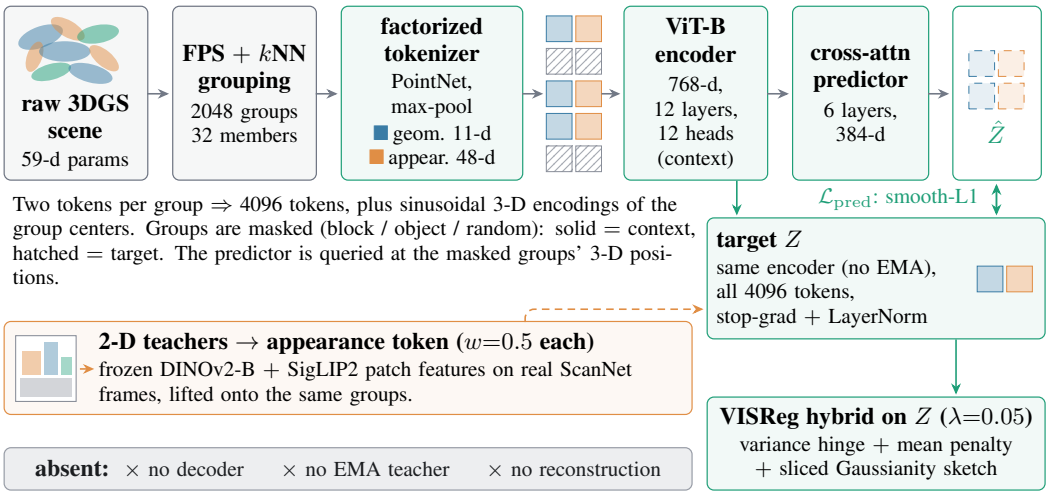


Figure 1: **Predicting masked embeddings from a Gaussian scene.** Local Gaussian groups become geometry/appearance token pairs. The context encoder sees visible groups; the predictor queries masked-group embeddings at supplied centers. Targets come from a full-scene pass through the same encoder, with stop-gradient and a distributional regularizer replacing a momentum teacher. Image features supervise only the appearance token directly during training; inference needs no photographs. Our main model, the reference, is shown; Table 1 distinguishes the tested variants.

predicts how a captured scene's representation changes under a specified object edit, supplied as the action.

SceneSplat [4] distills vision–language features into splats over 7K scenes and Gaussian-MAE [6] reconstructs masked object parameters; concurrent JEPAs use a momentum teacher on object-level splats [43] or scene-level point clouds [40], where a latent predictor already infers an absent object's semantics, or factorize their targets [44]. Rendering-supervised pretraining predicts Gaussians that are rendered and compared with images [7; 8], and GaussianDream++ [9] supervises robot-policy tokens with Gaussian reconstruction and future-state prediction before removing its renderer at deployment; we take fitted splats as input, and no loss of ours compares rendered images with pixels. On point clouds, Point-JEPA [47] predicts point-patch embeddings from context, Sonata and Concerto [31; 36] set the label-free probe standard, latent targets outperform reconstruction [33], and action-conditioned JEPAs plan toward commanded goals [46] (further related work in Appendix I).

Contributions.

1. **Predictive learning from Gaussian scenes** (Sections 2 and 3.1): we learn scene representations from raw Gaussian parameters by predicting masked regions from context, with a distributional regularizer in place of a momentum teacher; the predicted embeddings retrieve real scene fragments for hidden regions (Figure 2), with no photographs or rendering at inference.
2. **A recognition–prediction trade-off** (Section 3.2): controlled image-supervision sweeps show that stronger image supervision raises semantic recognition while making scene representations less predictable from context, a cost that recognition accuracy alone does not reveal. Lifted DINOv2 features recognize better than ours but are far less predictable under matched predictors, and routing image supervision to one token of each geometry/appearance pair keeps more predictability than distilling a single token.
3. **Spatially sensitive prediction** (Section 3.2): encoding relative spatial position in the predictor makes its predictions depend on how scene content is arranged, and turns prediction error into a signal for objects placed where they do not belong.
4. **Action-conditioned latent world modeling** (Figure 4, Section 3.3): a separately trained world model predicts how a captured scene's representation changes under synthesized

object edits, and searching over its predictions recovers the edits that produced a goal scene far more often than chance.

2 METHOD

Representing a scene. Prediction operates on groups of nearby Gaussians rather than on individual primitives. Farthest-point sampling selects group centers; a PointNet [61] pools each center's 32 nearest neighbors, expressed in relative coordinates, into a token. Inputs are the raw 59-dimensional position, log-scale, quaternion, opacity-logit and spherical-harmonic parameters. Sinusoidal encodings of group centers, standardized per scene, supply position separately from local content (details in Appendix B.1).

Learning to predict masked representations. A ViT-B encodes the visible groups into context embeddings Z_{vis}, and a cross-attention predictor Pred queries them at the supplied masked-group centers. Its targets are the embeddings Z of a second, full-scene pass through the *same* encoder. Layer normalization and stop-gradient define the regression target; a distributional regularizer receives gradients through the full-scene pass and prevents collapse without a momentum teacher:

$$\mathcal{L} = \text{smooth-L1}(\text{Pred}(Z_{\mathrm{vis}}), \mathrm{sg}[\mathrm{LN}(Z)_{\mathrm{masked}}]) + 0.05\,\mathcal{R}(Z) + \mathcal{L}_{\mathrm{aux}}.$$

$\mathcal{R}$ acts on the full-scene embeddings before normalization; the regression target is detached, while $\mathcal{R}$ and $\mathcal{L}_{\mathrm{aux}}$ train the encoder. Sharing the encoder removes the momentum copy, its update and its schedule; each step still encodes visible groups and the full scene. The base recipe uses our SIGReg implementation, a sliced Gaussianity regularizer following LeJEPA [11], and no auxiliary loss (exact objectives in Appendix B). Scene masks mix spatial blocks, random masks and object-aware masks, which use instance boundaries, not semantic classes (weights, proxy-mask controls and masked fractions in Appendix B.1); the supplied-center interface follows masked point prediction [47] (Limitation (5)).

Separating prediction from image supervision. Our main model, the *reference*, represents each group by a geometry/appearance token pair. Both participate in latent prediction; only the appearance token receives DINOv2 [53] and SigLIP2 [38] targets lifted from real photographs onto each group's visible members (a *routed* pair; Appendix B). This separates the direct image loss from one member of the pair, while shared attention still mixes information. The auxiliary loss is a symmetric InfoNCE between each predicted masked token and its full-scene target, with the batch's other masked tokens as negatives, plus one loss per image teacher. We use a scale–shape regularizer inspired by VISReg [13]: variance and mean terms control scale, and an Epps–Pulley statistic on standardized embeddings controls shape, in place of VISReg's sliced-Wasserstein term; *VISReg* below denotes this hybrid. The rotary model adds mixed-axis 3D RoPE [25] to predictor cross-attention, making relative query–context position enter attention directly. Table 1 separates these models' roles.

An action-conditioned latent world model. A separate four-layer transformer is an action-conditioned latent world model [10; 60; 17; 18]. Its state is the frozen representation Z of a captured scene and its action a specified object edit; it predicts the change $\hat{\Delta}$ toward Z', the encoder's re-encoding of the edited scene ($\hat{Z} = Z + \hat{\Delta}$; true change $\Delta = Z' - Z$), with squared error and SIGReg [21]. Remove deletes the object's groups, move translates their centers by a supplied 3D vector, and add inserts standalone object tokens; after a removal, loss and metrics use kept tokens only. Routed pairs are averaged per group, so the reference's 4096 encoder tokens yield 2048 state vectors.

3 EXPERIMENTS

Data, supervision and evaluation. The controlled corpus is 400 ScanNet [51] scenes fitted with 3DGS-MCMC [27]: 320 for pretraining gradients and 80 for validation. Semantic labels train only the frozen-feature probe, on 338 scenes with the 62 official validation scenes held out; instance boundaries define object masks and synthetic edits. Inductive corpora exclude every probe evaluation room and grow to 1381 ScanNet scans of 645 rooms; Table 6 lists each corpus's exposure to

Table 1: Configurations and their roles. G/T: groups/encoder tokens. All latent-prediction rows use the same encoder family; *image-free* excludes image teachers. Run identifiers and full settings are in Appendix B.1.

Model	G/T	Training recipe	Role
Base	$512/512$	JEPA, SIGReg, image-free	Objective comparison
Reconstruction	$512/512$	Parameters, with/without SIGReg	Control for base
Reference	$2048/4096$	JEPA, VISReg, routed teachers	Frontier; edit state
Rotary	$2048/4096$	Reference + predictor RoPE	Spatial prediction
Image-free pair	$2048/4096$	JEPA, VISReg, no image loss	Teacher control
Sibling	$1024/2048$	Reference at fewer groups	Secondary downstream space

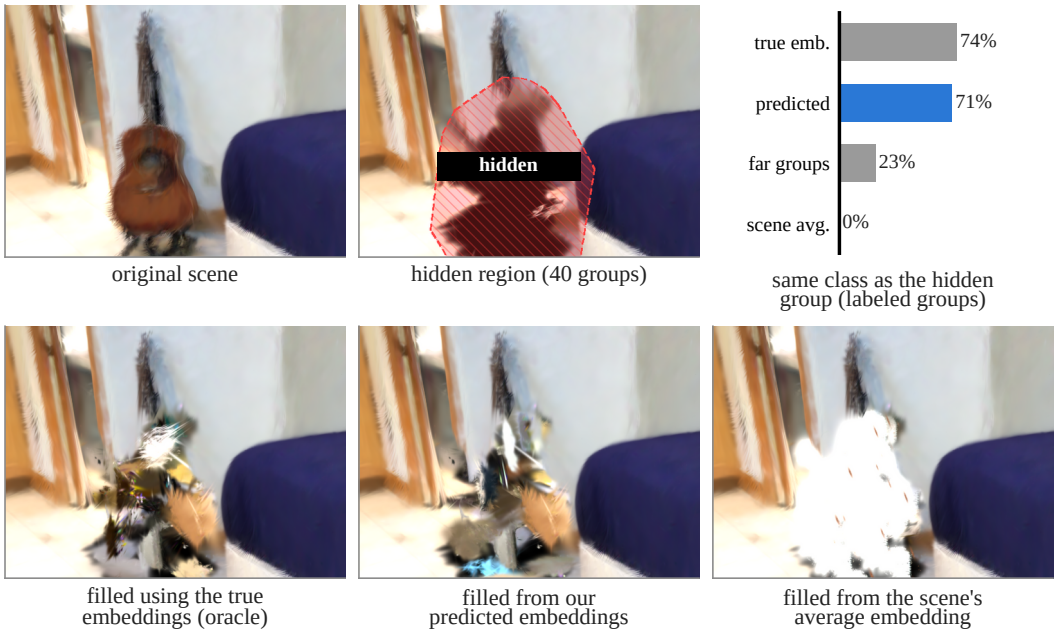


Figure 2: **Predicted embeddings retrieve hidden content almost as well as the true ones.** The rotary model predicts an embedding for each of 40 hidden groups around a guitar from the visible scene alone; each hidden group is then filled with the nearest of 305,152 real Gaussian groups from 149 other scenes, translated into place. Filling with the encoder's own embeddings of the hidden groups (a true-embedding oracle, unavailable at test time) looks just as fragmented: retrieval pastes real fragments rather than generating geometry. The scene's average embedding retrieves the same group for every hidden group. Bars: class agreement over the 35 labeled hidden groups (unlabeled targets excluded; far groups: random groups at least 1 m away). The region's majority class is floor. Population retrieval (reference encoder, all scenes): Table 15.

the separate prediction-evaluation scenes. Transfer uses InteriorGS [55; 56] and ModelNet40 [52] splats; real recapture uses 3RScan [29] and ScanNet (Section 3.4). Recognition is per-Gaussian voted probe mIoU. Predictability is the counterfactual (CF) cosine between predicted and full-scene embeddings at masked groups; because raw CF is difficulty-confounded (coarser groups sit closer to the scene mean), we report the prediction *margin*, CF minus a scene-mean null (Section 3.2). Recipes were selected on the official validation scenes we report; half-split resampling of those scenes puts the resulting probe optimism at 0.010 mIoU, and paired tests are descriptive without family-wise correction (Appendix C.3). Pre-registered gates are reported with their experiments.

3.1 LATENT PREDICTION VERSUS PARAMETER RECONSTRUCTION

We compare latent prediction and parameter reconstruction on recognition and predictability with a shared tokenizer and encoder; every regularized model must pass encoder-space collapse checks [32; 15; 16; 14]. Reconstruction follows Point-MAE [50] and Gaussian-MAE [6], and its regularized recipe adds two faithfulness fixes (Appendix D.3); we report it without a regularizer (*bare*, the baseline, which collapses dimensionally as U-MAE [30] predicts for plain masked autoencoders [63]) and with SIGReg [11].

Table 2: Latent prediction versus parameter reconstruction with matched tokenizer and encoder (regularized arms share λ): official-val probe mIoU, 5-way 10-shot accuracy on ModelNet40 test-split splats (100 fixed episodes shared by all arms; $\pm$ over model seeds, n=5/1/3; Appendix D.3), and CF prediction cosine in latent space.

	probe (official)	few-shot (5-way 10-shot)	counterfactual
latent JEPA (base recipe)	0.1059 ± 0.0018 (n=5)	$77.1 \pm 6.2\%$	$\mathbf{0.848}$
recon, bare (baseline)	0.110 (collapsed)	35.5%	≈ 0
recon + SIGReg	$\mathbf{0.137}$–$\mathbf{0.151}$ (noncollapsed)	$\mathbf{91.1 \pm 0.4}\%$	≈ 0

Without a regularizer, reconstruction collapses and ties the latent model on the probe; with SIGReg it passes collapse checks and probes higher (Table 2). On few-shot object recognition the probe-tied pair separates sharply: latent prediction more than doubles the accuracy of bare reconstruction (whose checkpoint predates both fixes), and regularized reconstruction is higher still, so similar scene probes can conceal large differences in object recognition. On CF prediction the latent model scores 0.848 (slightly lower when trained and evaluated on InteriorGS; Appendix D.1) and every reconstruction variant ≈ 0 in latent space; scored in its own output space, regularized reconstruction beats scene-mean and random substitutes only narrowly (Appendix D.3). CF depends on the trained prediction interface (Appendix D.3; cf. Point2Vec and Sonata [33; 31]), but a matched fresh predictor on the frozen encoders of Table 3 shows that regularized-reconstruction features are far less predictable from visible context than JEPA features (CF 0.562 vs. 0.877; three predictor seeds; Table 10).

Table 3: The 2×2: JEPA loss $\times$ distillation on one 260-scene split with real-frame targets and one token layout (seed 42; last row n=3). The distillation-only recipe also has no regularizer, so rows differ as recipes, not by one factor.

Configuration	probe (mIoU)	counterfactual (cosine)
neither: reconstruction + SIGReg	0.143	0.000
JEPA only	0.115	0.849
distillation only (λ_{JEPA}=0; predictor untrained)	0.249	-0.001
both	0.277	0.698
both + per-token InfoNCE (n=3)	$\mathbf{0.307 \pm 0.008}$	$\mathbf{0.895}$

Separating predictive learning from image distillation. We compare JEPA and image-distillation recipes, then add the reference's per-token contrastive term to the joint recipe (Table 3). Distillation alone doubles the probe; its near-zero CF reflects a predictor that no loss trains, and fresh predictors on the frozen encoders' visible groups show that features trained with distillation, alone or with the JEPA term, are much less predictable from visible context (CF 0.675–0.694 vs. 0.877 for JEPA only; three predictor seeds; Table 10). Trained end to end, adding distillation raises the probe and lowers CF, and the per-token contrastive term restores CF while keeping the probe gain. Appendix B.1 covers masking and regularizers, Appendix C.2 the oracle ceiling, and Section 1 of the Extended results (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`) the input-feature ablation.

3.2 RECOGNITION VERSUS PREDICTION

Image supervision trades recognition for prediction. Raising the image-teacher weight on the single-token recipe (Figure 3) more than quadruples the probe while the scene-mean and random-

Table 4: Main results on official ScanNet-20 val (transductive probes: five encoder seeds for the base recipe, the 2048-token VISReg-inspired pair, the reference, the 8192-token pair and the rotary model, three elsewhere, including every inductive entry; s.d. in Appendix D). Probe: voted linear mIoU, transductive / inductive (evaluation scenes inside / outside the pretraining corpus). CF: counterfactual prediction cosine (main sweep). Margin: CF in the null-control run (one instance per scene) minus its scene-mean null (Appendix D.5). *Inductive* qualifies the probe column only. CF and margin use a separate 20-scene set whose training exposure depends on the corpus; the 1381-scan models saw 17 of those scenes in pretraining; their CF values come from the null-control pass.

model	tokens	probe trans. / ind.	CF	margin
oracle partition ceiling (majority label per token)	2048	0.928 / —	—	—
teacher features lifted from photos, no JEPA (one run)	—	0.654 / —	—	—
teacher features lifted from splat renders, no JEPA (one run)	—	0.481 / —	—	—
SceneSplat checkpoint, our probe (VL-supervised, 7K scenes; one run)	—	0.651 / —	Tab. 12	
base recipe (no image teacher)	512	0.106 / 0.115	0.848	—
token pairs, VISReg, no image teacher	4096	0.132 / —	**0.945**	**0.940**
no image teacher, self-distillation on the encoder output, 1381-scan corpus	4096	— / 0.188	0.922	0.930
routed pairs, VISReg	2048	0.507 / 0.283	0.892	0.816
both tokens distilled, SIGReg	2048	**0.541** / 0.316	0.891	0.776
routed pairs, VISReg (reference)	4096	0.529 / 0.286	0.886	0.830
reference + rotary position on the predictor	4096	**0.534** / —	**0.889**	**0.833**
routed pairs, VISReg, k=16, 200k cap	8192	**0.551** / **0.313**	0.883	0.840
single token, VISReg, k=16, 200k cap	4096	**0.565** / —	0.869	0.789
routed pairs, VISReg, inductive, 1056-scan corpus	2048	— / 0.410	0.900	0.847
reference recipe, inductive, 1056-scan corpus	4096	— / 0.433	0.898	0.853
reference recipe, inductive, 1381-scan corpus	4096	— / 0.448	0.904	0.857
rotary recipe, inductive, 1381-scan corpus	4096	— / **0.460**	**0.905**	**0.861**

token margins fall monotonically; past the default weight the probe saturates while the margins keep falling, and a shuffled-teacher control keeps only 18% of the probe gain (Appendix D.4). The dissociation holds across representation families: with matched predictors on cached full-scene features, lifted DINOv2 features, which out-probe ours, reach a scene-mean margin of 0.56 against 0.93 for our frozen encoder under the same protocol (Table 12). The image-free pair recipe probes near the base recipe at the largest margin in Table 4, and more data does not substitute for images: without evaluation rooms, adding scenes leaves the 2048-token image-free pair recipe's probe at 0.13–0.14 but lifts the reference recipe's from 0.286 to 0.448 (n=3 each; Appendix C.1).

How the reference model is built. Matched comparisons favor real-frame over rendered-view teacher targets [36], routed token pairs for scene-mean margin, and our VISReg-inspired hybrid over SIGReg for margin in a test paired by scene, at a similar probe (Appendix D.4; Figure 6 shows all models). At 2048 tokens pairs trade some probe for margin, and they raise margins over nulls that ignore context but not over context-sensitive ones (Tables 12 and 11). At 4096 tokens the reference reaches CF 0.886 (n=5), just short of the pre-registered CF $\geq$ 0.90 bar and above the highest-probe model (Table 4).

What the prediction margin measures. Because the scene-mean null falls with token count, margins compare models at one token count, and shuffled- and mean-context nulls show that the reference predictor uses the content of its context, not its arrangement (Appendix D.5). These nulls probe the masked-region predictor rather than the encoder whose full-scene output the world model reads (Section 3.3).

Predicted embeddings retrieve hidden content. Filling each hidden group with the real group from other scenes nearest to its predicted embedding matches the hidden group's class almost as often as filling it from the true embedding (Figure 2). Across scenes and with matched predictors, the reference and standardized SceneSplat spaces reach 99% and 98% of their true-embedding oracles

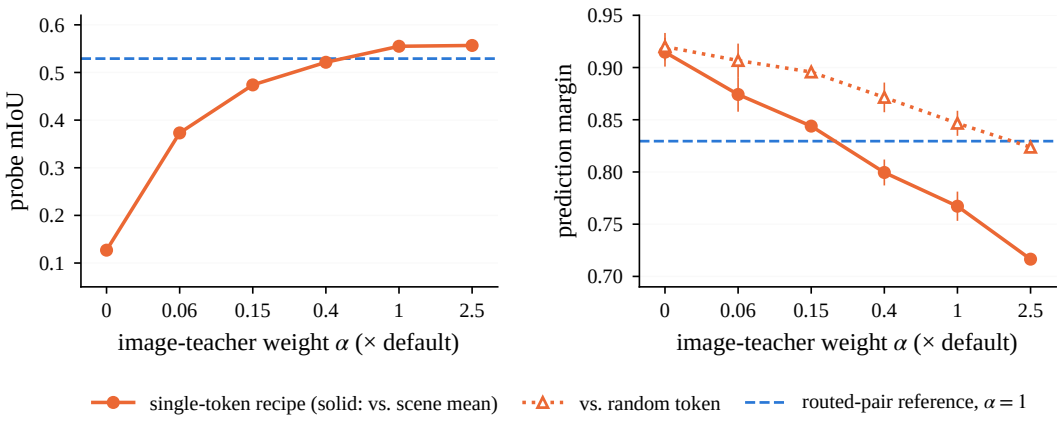


Figure 3: **Image supervision trades recognition for predictability.** One single-token recipe with its image-teacher weight α scaled from 0 (no teacher) to $2.5\times$ the default: the probe rises and saturates while both prediction margins fall monotonically. Dashed: the routed-pair reference at the default weight (left: probe; right: scene-mean margin), which keeps a higher margin at a slightly lower probe. The six weights are discrete settings, spaced evenly. Sample s.d.; n=2–3 per weight, n=5 for the reference.

in group-level class agreement, and both beat copying the nearest visible group: retrieval follows the space's class structure (Appendix F).

Rotary position makes prediction spatial. Adding 3D rotary position embeddings [25] to the predictor's cross-attention raises the shuffled-context margin from near zero to 0.63 (n=5) and nearly triples the mean-context margin, with similar probe, CF and scene-mean margin (Table 13); the pattern repeats on two larger corpora (Appendix E). Over all instances rather than the largest per scene, the rotary predictor clearly beats a copy of the nearest visible token's context embedding (CF 0.87 vs. 0.65), which the reference beats on objects above ∼40 groups but not over all instances (Figure 7). Its prediction error also flags transplanted objects resting on the floor, where height gives no signal, better than every tested static score in its own latent space (AUROC 0.88 vs. 0.80, three seeds; edited scenes re-encoded as in deployment); frozen-feature predictors score higher only under a protocol that flags any encoding mismatch (Table 14).

Generalization to unseen rooms. With no evaluation room in pretraining, the rotary recipe reaches **0.460** on 1381 scans (n=3; Table 4) and the reference recipe 0.448, closing two-thirds of the gap between the reference's controlled-scale inductive and transductive probes (on exposure, see Limitation (1)). On 80 validation rooms that the 400-scene models never saw, the rotary model probes higher than the reference on all five paired seeds (Table 9).

3.3 An action-conditioned latent world model

The world model predicts the representation a scene will have after a specified object edit *in 3D*, without re-encoding the edited scene; re-encoding computes this *latent* target exactly and serves as an oracle for every task here. Its primary space is the reference's. Comparisons between spaces use synthesized transitions from 50 training and 15 validation ScanNet scenes, with 30 held-out scenes for composition and search. The sibling and base-recipe spaces are in Appendix G. **Remove** reaches relative L2 0.489 ± 0.017 (three world-model seeds), the mean of $\|\hat{\Delta}_i - \Delta_i\|/\|\Delta_i\|$ over the quarter of kept tokens that change most (Appendix B); identity scores 1, and non-learned transition baselines do no better (Appendix G). The predicted change (Figure 4) has direction cosine 0.861 ± 0.010 and beats identity on 98% of the disturbed tokens; the re-encoding targets include substantial coordinate-frame effects (Limitation (4)). With more training scenes, removal keeps improving: 150 and 385 scenes lower relative L2 to 0.369 and 0.277, and larger sets beat the 50-scene model even at earlier checkpoints with fewer gradient steps (Table 18); all other world-model results use the 50-scene set. Sweeping the **move** distance gives a psychometric detection threshold that depends on corpus

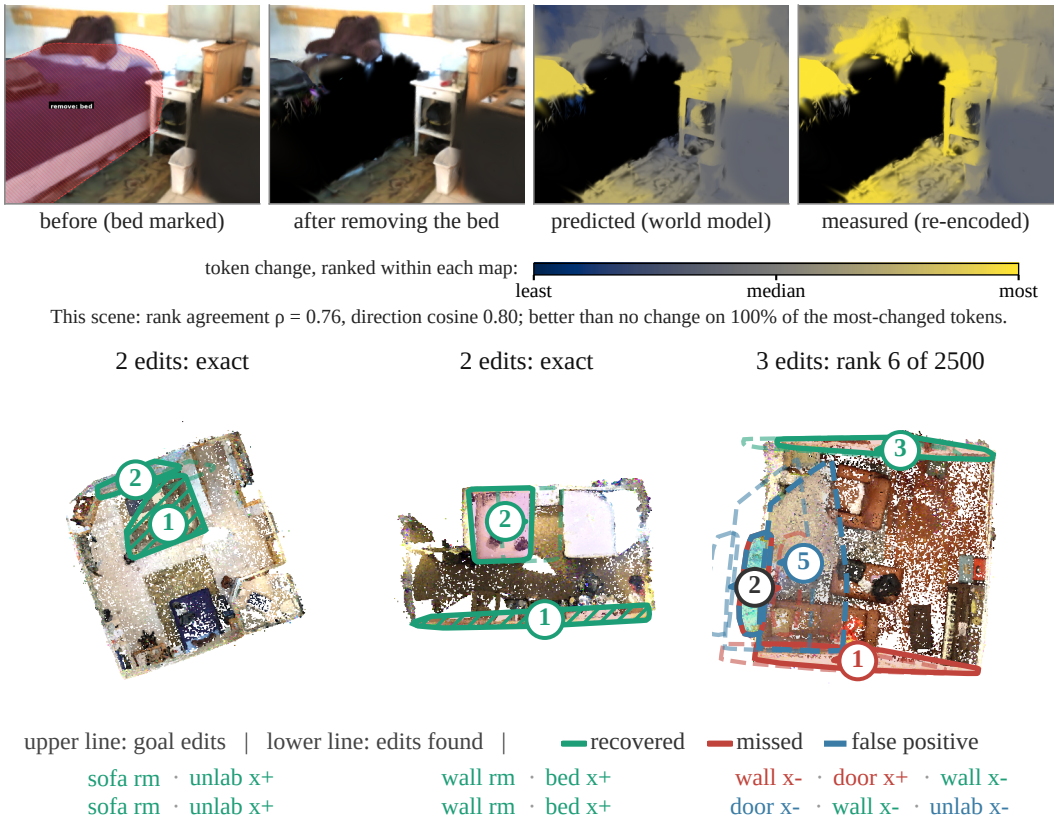


Figure 4: **The world model predicts how an edit changes a scene's representation, and searching over its predictions recovers edits.** Top: removing the bed (reference model). Colors rank each token's change within its own map, predicted and measured separately, so the maps compare where the scene changes rather than how much; the shared-scale maps (Figure 9) show that the prediction under-shoots magnitude. Part of the measured change is the re-normalized coordinate frame (Limitation (4)). Bottom: three held-out scenes seen from above (their own Gaussians). Upper line: the edits that produced the goal scene; lower line: the edits the search found (green recovered, red missed, blue false positive, black: right object, wrong action; rm: remove, x±: move along $\pm x$, unlab: unlabeled instance; hatch: remove; arrow: move, drawn capped). Two scenes are recovered exactly; in the third the true set ranks sixth of 2500.

and encoder (Appendix G). **Add** [26], the hardest action in our tests, still beats identity on 72% of context tokens, and two-step **composition** beats identity on all 30 held-out scenes (Table 16).

Edit search in latent space. Given a held-out scene and a goal scene made from it by two edits (object, remove vs. move, and move direction), scoring every candidate edit set by the world model's predicted end state recovers the true set in 31% of 90 trials, against 0.27% chance (three world-model seeds), and more often than a greedy heuristic that reads the goal's removed slots (exact McNemar $p{=}6 \times 10^{-5}$, first seed; Appendix G); executing the recovered edits cuts the median latent distance to the goal to a fifth, whereas random edits increase it. At three edits exact recovery is 14% against 0.04% chance ($p < 0.003$ at every seed). An exact-simulator oracle recovers every set at both horizons, so the remaining gap is world-model prediction error, not task ambiguity (Figure 8). In the rotary space the same protocol is harder: the two spaces favor different prediction tasks (Table 16).

3.4 Separating scene change from reconstruction variability

Evaluation on real rescan pairs separates reconstruction variability from the scene changes a world model should predict. A fit-noise floor appears in every encoder we tested: two *independently fitted*

reconstructions of the same unchanged region differ by a latent drift as large as most edit-induced change in the base-recipe and two real-frame-distilled encoders, and by most of the random-token distance in the reference encoder (Table 19). On Remove360 (base-recipe space), **frozen local refits** [35] of only the Gaussians near the change lower this floor from about 0.14 to at most 0.003, whereas independent and warm-initialized refits keep it; on real ScanNet rescans, frozen refits do not yet meet a pre-registered fidelity gate (Appendix H). A base-recipe world model trained only on synthesized edits did not transfer zero-shot to real pairs: it beat identity on 11 of 25 3RScan pairs without localizing the edit, and on none of the three frozen-refit scenes. Two measurements point to why: a frozen-refit removal changes the rest of a real scene little, while the synthetically trained model predicts larger disturbances on these real inputs than on its validation data. On synthesized transitions such models beat identity (Table 16), and frozen refits make real transitions measurable to train on.

4 Limitations and next steps

(1) Exposure. Controlled-corpus probes are transductive because pretraining includes the 62 evaluation scenes' splats and photo-lifted targets (never their class labels); excluding those scenes keeps 54–69% of each distilled probe while the reference's margin stays similar; a target-removal control traces the difference to the photographs rather than the splats; and the inductive corpora of Table 4 avoid this exposure (Appendix C.1). **(2) Seed counts.** Encoder seeds range from one to five (five for the controlled-corpus reference and rotary rows); every value states its n, and Appendix C.3 gives the detectable differences. **(3) Synthesized edits.** The world model learns from synthesized edits (moves translate without rotating; removals leave the occluded surface unsynthesized) because the fit-noise floor hides real change (Section 3.4). Frozen local refits remove that floor; once they pass the pre-registered fidelity gate on real rescans, real refit transitions become the training source and the test of real-edit transfer (Appendix H). **(4) Coordinate normalization.** Centers are standardized per input, so an edit also shifts the coordinate frame, which dominates the change of the remaining tokens (Table 17); trained on fixed-frame targets in a 200-epoch follow-up, the world model clears the pre-registered fixed-frame alignment threshold in all five seeds (Appendix G). **(5) Supplied geometry.** As in masked point prediction, the predictor is queried at the masked groups' own centers, so a target's footprint is given and overlapping k-NN groups can leave part of a masked instance visible; the mean-context null and the nearest-visible copy rule out two simple uses of the footprint but not every use; querying without supplied centers would remove this cue (Appendix D.5). Appendix A.1 gives each item in full.

5 Conclusion

We showed that scene representations can be learned from raw Gaussian parameters by predicting masked regions in latent space, without photographs at inference. Measuring recognition and prediction separately reveals a trade-off that recognition alone misses: image supervision makes content easier to recognize but harder to predict from context. Relative position in the predictor makes prediction depend on scene arrangement. These representations fill hidden regions with real Gaussian groups and support an action-conditioned latent world model, trained on synthesized edits, whose removal predictions improve steadily with more training scenes and which is accurate enough to search for the edits that produced a goal scene. Fixed-frame training captures removal effects beyond coordinate normalization, and frozen local refits remove the fit-noise floor that hides real scene change; once they reconstruct real rescans faithfully, training on real transitions is the natural next step. Together, the results show that captured splats support both inference of hidden content and action-conditioned prediction of scene representations.

AI use statement

AI assistants supported this work: they helped write and debug code, find related work, suggest control experiments and analyses, and draft and edit text. The authors defined the research questions, directed and checked every experiment, verified all results and citations, and take full responsibility for the paper.

Reproducibility statement

Appendix B gives the exact objectives and Appendix C the splits, exposure, statistical units and compute. Code, the saved configurations of the reported models and per-scene evidence for the main controls are at `https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`. Unless a table says otherwise, $\pm$ is the sample standard deviation over encoder seeds, and every result states its seed count.

Ethics statement

We use ScanNet, 3RScan, ModelNet40, InteriorGS, Remove360 and CMU Panoptic Studio under their research licenses and redistribute none of them. We collected no data from people; the PanopticSports sequences show performers recorded by the data provider. We see no risks specific to this work beyond the dual-use considerations common to 3D scene understanding.

References

[1] Y. Qiu, R. Zhu, Y.-C. Chen. Improving Joint Embedding Predictive Architecture with Diffusion Noise (N-JEPA). arXiv:2507.15216.

[2] X. Chen, Z. Liu, S. Xie, K. He. Deconstructing denoising diffusion models for self-supervised learning. arXiv:2401.14404.

[3] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics 42(4) (SIGGRAPH), 2023.

[4] Y. Li et al. SceneSplat: Gaussian Splatting-based Scene Understanding with Vision-Language Pretraining. ICCV 2025. arXiv:2503.18052.

[5] M. Ma et al. SceneSplat++: A Large Dataset and Comprehensive Benchmark for Language Gaussian Splatting. arXiv:2506.08710.

[6] Q. Ma et al. ShapeSplat: A Large-scale Dataset of Gaussian Splats and Their Self-Supervised Pretraining. 3DV 2025. arXiv:2408.10906.

[7] H. Liu, M. Chen, Y. Ma, H. Xiao, Y. He. Point Cloud Unsupervised Pre-training via 3D Gaussian Splatting. arXiv:2411.18667, 2024.

[8] J. Rajasegaran, X. Chen, R. Li, C. Feichtenhofer, J. Malik, S. Ginosar. Gaussian Masked Autoencoders. arXiv:2501.03229, 2025.

[9] Y. Jiang, Z. Zhang, W. Zhou, J. Wang, J. He, L. Yang, H. Qing, S. Liu, D. Zhao, P. Luo, H. Yu. GaussianDream++: Efficient 3D Gaussian World Modeling for Robotic Manipulation. arXiv:2608.25659, 2026.

[10] Y. LeCun. A Path Towards Autonomous Machine Intelligence. OpenReview BZ5a1r-kVsf, 2022.

[11] R. Balestriero, Y. LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics. arXiv:2511.08544.

[12] A. Bardes, J. Ponce, Y. LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. ICLR 2022.

[13] H. Wu, R. Balestriero, M. Levine. VISReg: Variance-Invariance-Sketching Regularization for JEPA training. arXiv:2606.02572.

[14] R. J. Alvarez. Beyond Isotropy in JEPAs: Hamiltonian Geometry and Symplectic Prediction (HamJEPA). arXiv:2605.20107.

[15] K. Agrawal et al. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. NeurIPS 2022 (proceedings; OpenReview ii9X4vtZGTZ).

[16] V. Thilak et al. LiDAR: Sensing Linear Probing Performance in Joint Embedding SSL Architectures. arXiv:2312.04000.

[17] J. Richens, T. Everitt. Robust Agents Learn Causal World Models. ICLR 2024. arXiv:2402.10877.

[18] B. Schölkopf et al. Towards Causal Representation Learning. Proc. IEEE 2021. arXiv:2102.11107.

[19] Q. Garrido et al. Learning and Leveraging World Models in Visual Representation Learning (IWM). ICML 2024. arXiv:2403.00504.

[20] G. Zhou et al. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. arXiv:2411.04983.

[21] L. Maes, Q. Le Lidec, D. Scieur, Y. LeCun, R. Balestriero. LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels. arXiv:2603.19312.

[22] J. U. Kreber, L. Mack, J. Stueckler. Learning Action-Conditional and Object-Centric Gaussian Splatting World Models for Rigid Objects (MRO-GWM). arXiv:2606.01950.

[23] G. Lu et al. GWM: Towards Scalable Gaussian World Models for Robotic Manipulation. ICCV 2025. arXiv:2508.17600.

[24] G. Lu et al. ManiGaussian: Dynamic Gaussian Splatting for Multi-task Robotic Manipulation. arXiv:2403.08321.

[25] B. Heo, S. Park, D. Han, S. Yun. Rotary position embedding for vision transformer. ECCV, 2024. arXiv:2403.13298.

[26] D. Winter et al. ObjectDrop: Bootstrapping Counterfactuals for Photorealistic Object Removal and Insertion. ECCV 2024. arXiv:2403.18818.

[27] S. Kheradmand et al. 3D Gaussian Splatting as Markov Chain Monte Carlo. NeurIPS 2024. arXiv:2404.09591.

[28] S. Kocour et al. Remove360: Benchmarking Residuals After Object Removal in 3D Gaussian Splatting. arXiv:2508.11431.

[29] J. Wald et al. RIO: 3D Object Instance Re-Localization in Changing Indoor Environments (3RScan). ICCV 2019. arXiv:1908.06109.

[30] Q. Zhang, Y. Wang, Y. Wang. How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders (U-MAE). NeurIPS 2022. arXiv:2210.08344.

[31] X. Wu et al. Sonata: Self-Supervised Learning of Reliable Point Representations. CVPR 2025. arXiv:2503.16429.

[32] Q. Garrido, R. Balestriero et al. RankMe: Assessing the Downstream Performance of Pretrained Self-Supervised Representations by Their Rank. ICML 2023. arXiv:2210.02885.

[33] K. Knaebel, J. Schult, A. Hermans, B. Leibe. Point2Vec for Self-Supervised Representation Learning on Point Clouds. GCPR 2023. arXiv:2303.16570.

[34] J. Luiten et al. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. 3DV 2024. arXiv:2308.09713.

[35] J. Ackermann et al. CL-Splats: Continual Learning of Gaussian Splatting with Local Optimization. arXiv:2506.21117.

[36] Y. Zhang, X. Wu, Y. Lao, C. Wang, Z. Tian, N. Wang, H. Zhao. Concerto: Joint 2D-3D Self-Supervised Learning Emerges Spatial Representations. NeurIPS 2025. arXiv:2510.23607.

[37] F. Bordes, R. Balestriero, Q. Garrido, A. Bardes, P. Vincent. Guillotine Regularization: Why removing layers is needed to improve generalization in self-supervised learning. TMLR 2023. arXiv:2206.13378.

[38] M. Tschannen et al. SigLIP 2: Multilingual vision–language encoders with improved semantic understanding, localization, and dense features. arXiv:2502.14786.

[39] L. Landrieu, M. Boussaha. Point Cloud Oversegmentation with Graph-Structured Deep Metric Learning. CVPR 2019. arXiv:1904.02113.

[40] Z. Zhou, Q. Wen, X. Zeng. SR-JEPA: Learning Predictive Latent State in 3D Scenes. arXiv:2608.05774.

[41] J. Yeom, H. Kim, J. Park, S. Jung, J. Lee, T. Kim. What Makes Video World Model Latents Action-Relevant: Prediction over Reconstruction. arXiv:2606.07687.

[42] Y. Huang. 4DGS-JEPA: Temporally compositional joint-embedding prediction for dynamic Gaussian splatting. arXiv:2609.25036, 2026.

[43] B. Ren, Q. Ma, Y. Li, Z. Han, Y. Li, Y. Fu, R. M. Anwer, T. Gevers, F. S. Khan, S. Khan. Gaussian-JEPA: Joint-Embedding Predictive Learning for 3D Gaussian Splats. arXiv:2608.15651.

[44] T. Cui, Z. Wang, X. Xu, W. Liu, Z. Yu, Y. Zhang, Q. Gao, M. Yang, W. Ouyang, P. A. Heng, Y. Wu, Z. Yin, L. Yang. JEPA-Anything: Learning Predictive Models across Different Worlds. arXiv:2609.20800, 2026.

[45] Y. Ma, Z. Xu, I. King. 4DGS-WAM: Bridging Past and Future with an Object-Centric World Action Model based on 4D Gaussian Splatting. arXiv:2608.25956, 2026.

[46] F. F. Oberweger, M. Schwingshackl. Does Latent Planning Survive Point Clouds? Action-Conditioned JEPA World Models for Geometric Observations. arXiv:2608.29434, 2026.

[47] A. Saito, P. Kudeshia, J. Poovvancheri. Point-JEPA: A Joint Embedding Predictive Architecture for Self-Supervised Learning on Point Cloud. arXiv:2404.16432, 2024.

[48] Z. An, Y. Xiao, Z. Leng et al. Cross-temporal 3D Gaussian splatting for sparse-view guided scene update. AAAI, 2026. arXiv:2512.00534.

[49] T. Lentsch et al. Emergent 3D instance segmentation from self-supervised point transformers. arXiv:2608.15796, 2026.

[50] Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, L. Yuan. Masked Autoencoders for Point Cloud Self-supervised Learning. ECCV, 2022.

[51] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. CVPR, 2017.

[52] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes. CVPR, 2015.

[53] M. Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. TMLR, 2024. arXiv:2304.07193.

[54] H. Joo et al. Panoptic Studio: A Massively Multiview System for Social Motion Capture. ICCV, 2015.

[55] B. Miao, R. Wei, Z. Ge, X. Sun, S. Gao, J. Zhu, R. Wang, S. Tang, J. Xiao, R. Tang, J. Li. Towards Physically Executable 3D Gaussian for Embodied Navigation. ICLR 2026 (OpenReview HB6KvsqcAn).

[56] SpatialVerse Research Team, Manycore Tech Inc. InteriorGS: A 3D Gaussian Splatting Dataset of Semantically Labeled Indoor Scenes. 2025. https://huggingface.co/datasets/spatialverse/InteriorGS.

[57] J. Farganus, K. Żurawicki, A. Gaweł, W. Jakubowska, H. Kwaśnicka. Learning Representations from 3D Gaussian Splats. arXiv:2605.29549, 2026.

[58] M. Assran et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. CVPR 2023. arXiv:2301.08243.

[59] A. Bardes et al. Revisiting Feature Prediction for Learning Visual Representations from Video. TMLR 2024. arXiv:2404.08471.

[60] M. Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv:2506.09985, 2025.

[61] C. R. Qi, H. Su, K. Mo, L. J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017. arXiv:1612.00593.

[62] X. Wu et al. Point Transformer V3: Simpler, Faster, Stronger. CVPR 2024. arXiv:2312.10035.

[63] K. He et al. Masked Autoencoders Are Scalable Vision Learners. CVPR 2022. arXiv:2111.06377.

[64] A. van den Oord, Y. Li, O. Vinyals. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748, 2018.

A Overview

Appendix B states the exact objectives; Appendix C the splits, exposure, statistics and compute; Appendix D the recognition and prediction results with their matched controls and nulls; Appendices E–H spatial prediction, retrieval, the latent world model and the transfer boundary; Appendix I related work. The *Extended results* document in the anonymous repository (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`) gives the unabridged text of every section, secondary ablations and further figures; its numbered sections are cited below.

A.1 Limitations in detail

(1) Exposure. Controlled-corpus probes are transductive: the encoder saw the 62 evaluation scenes, without class labels (label-free proxy masks: Appendix B.1). Excluding them lowers every distilled model's probe to 54–69% of its transductive value, while the image-free base recipe's probe and the reference's margin stay similar (Figure 5); keeping those rooms' splats but removing their photo-lifted targets brings the probe level with the inductive one (Appendix C.1): in this control the transductive gain comes from the rooms' photographs, not their splats. The inductive corpora exclude every evaluation room, and Table 4 reports their probes beside the transductive ones.

(2) Seed counts. Controlled-corpus reference and rotary rows use five encoder seeds, larger-corpus rows three and some controls one; frozen-feature comparisons use three predictor seeds on one encoder (units and detectable differences: Appendix C.3).

(3) Synthesized, translation-only edits. The world model learns from transitions synthesized from annotated scenes; move translates an object without rotating it, and removal does not synthesize the surface the object occluded. Real-edit transfer is untested in the reference space. Two independently fitted reconstructions of one unchanged region differ by a large latent drift, the fit-noise floor (Section 3.4, Table 19); frozen local refits remove this floor on Remove360, but frozen refits of real ScanNet rescans do not yet meet the pre-registered fidelity gate (Appendix H), so real refit transitions are the next training source.

(4) Coordinate normalization. Centers are standardized per input (Appendix B), so an edit also changes the coordinate frame; this dominates the change of the remaining tokens, whereas the moved object's change aligns more with the fixed-frame edit (Table 17). Trained on standard targets, the world model's projected fixed-frame alignment is 0.111 (below the pre-registered 0.30, yet positive in all 15 scenes); trained on fixed-frame targets for 200 epochs in a follow-up, it exceeds that threshold in all five seeds (Appendix G).

(5) Supplied geometry. The predictor is queried at the masked groups' own centers, so a target's footprint is given even when its content is not, and overlapping k-NN groups can leave part of a masked instance visible; beating the mean-context null and the nearest-visible copy (for the reference, on large objects only) rules out those predictors but does not bound how a learned predictor uses the footprint (Appendix D.5).

B Objectives and implementation

This section states the implemented objectives exactly; code paths refer to the supplementary code (`models/sigreg.py`, `models/jepa.py`, `models/world_model.py`).

Pretraining loss. Let $Z \in \mathbb{R}^{n\times d}$ (d=768) collect the full-scene encoder outputs of every token in a batch before layer normalization (both tokens of each routed pair; n = batch size $\times$ tokens per scene), and let M index the masked tokens. The loss is

$$\mathcal{L} = \text{smooth-L1}_{\beta=2}\big(\,\text{Pred}(Z_{\text{vis}})_M,\ \text{sg}[\text{LN}(Z)]_M\big) + \lambda\,\mathcal{R}(Z) + \mathcal{L}_{\text{aux}}, \qquad \lambda = 0.05.$$

The regression target is detached; $\mathcal{R}$ and $\mathcal{L}_{\text{aux}}$ are not, so they train the shared encoder through the full-scene pass. The reference uses batches of 4 scenes of 4096 tokens; the base recipe, batches of 6 scenes of 512 tokens.

The VISReg-inspired hybrid (reference, rotary, image-free pair and every model labeled VISReg). With column means μ_j, sample standard deviations $s_j = \sqrt{\mathrm{Var}_i(Z_{ij}) + 10^{-4}}$ and standardized rows $\tilde{z}_i = (z_i - \mu)/s$,

$$\mathcal{R}_{\mathrm{VIS}}(Z) = \frac{1}{d}\sum_j \max(0, 1 - s_j) + \frac{1}{d}\sum_j \mu_j^2 + \frac{1}{m}\sum_{r=1}^{m}\sum_{k=0}^{16} w_k \Big| \frac{1}{n}\sum_i e^{\,\mathrm{i}t_k u_r^\top \tilde{z}_i} - e^{-t_k^2/2} \Big|^2,$$

with m=1024 unit directions u_r (normalized Gaussian draws, redrawn at every training step), $t_k = 3k/16$, and trapezoid weights $w_k = 2\Delta t\, e^{-t_k^2/2}$ ($\Delta t = 3/16$), halved at $k = 0$ and $k = 16$. The first two terms control scale and location; the third is the Epps–Pulley statistic of LeJEPA [11] against the fixed $\mathcal{N}(0, 1)$ characteristic function, applied to standardized embeddings so that it judges shape only. VISReg [13] also separates scale from shape, but its shape term is a sliced-Wasserstein distance; we did not run that objective, and no comparison between the two is claimed.

Our SIGReg implementation (base recipe and every model labeled SIGReg). With projections $p_{ir} = u_r^\top z_i$ of the raw embeddings, per-direction means a_r and variances v_r, and the same m=1024 resampled directions,

$$\mathcal{R}_{\mathrm{SIG}}(Z) = \frac{1}{m}\sum_r \left[a_r^2 + (v_r - 1)^2\right] + \frac{1}{4}\sum_{t\in\{0.5,1,1.5,2\}} \frac{1}{m}\sum_r \Big| \frac{1}{n}\sum_i e^{\,\mathrm{i}tp_{ir}} - e^{\,\mathrm{i}ta_r - t^2 v_r/2} \Big|^2.$$

The first term standardizes every projection; the second compares each projection's empirical characteristic function with that of a Gaussian of the same mean and variance, so it constrains shape. This is a sliced Gaussianity regularizer in LeJEPA's family, not LeJEPA's exact fixed-target statistic. The world model uses the same form on its predicted states (kept tokens, 128 directions, weight 0.05).

Token InfoNCE. A symmetric InfoNCE [64] (temperature 0.07, weight 0.1 in $\mathcal{L}_{\mathrm{aux}}$) contrasts each ℓ_2-normalized predicted masked token with its target $\mathrm{LN}(Z)_i$ against every other masked token of the batch, the other token of the same group included; geometry and appearance tokens are never positives for each other. Unlike the regression target, its target side is not detached, so it reaches the full-scene representation.

Image-teacher targets. Features are `dinov2_vitb14` patch tokens (518×392 input, patch 14) and `google/siglip2-base-patch16-512` [38] patch tokens (512×512), both from real ScanNet color frames. Up to 60 posed frames per scene, evenly spaced, are used. Each member Gaussian of the frozen training grouping is projected into every frame; an offline rasterization of the fitted scene at the same pose supplies accumulated alpha α and expected depth D, and a member at depth z receives the soft visibility weight

$$w = \Big(1 - \alpha\,\sigma\big(\tfrac{z-D}{s}\big)\Big)\,\sigma\big(\tfrac{\alpha-0.1}{0.1}\big), \qquad s = 0.05D + 0.10,$$

so the renderer is used offline for visibility only, never as a training loss or at inference. Members with total weight below 10^{-3} fall back to their two nearest in-frustum cameras. A group's target is the average over its covered members; groups with no covered member receive a zero target and are excluded from the loss. The student side is a two-layer MLP head per teacher (768→768→768, GELU) on the appearance token's final embedding, trained with a cosine loss ($1 - \cos$), weight 0.5 each. The target-removal control zeroes the targets of every capture of the 62 evaluation rooms and changes nothing else.

Coordinates. Group centers are standardized per encoder call: subtract their mean and divide each axis by its standard deviation over all groups of that input (clamped at 10^{-6}). Member coordinates stay in scene units relative to their group center. Because the statistics are recomputed on every call, re-encoding after a removal or a move re-derives the frame, which shifts the normalized position of every token, not only the edited object's (Limitation (4)).

World-model transitions and metrics. Let $[N]$ index a scene's $N{=}2048$ state vectors (routed pairs averaged), O the edited object's groups and $K = [N] \setminus O$. *Remove:* the kept groups are re-encoded and scattered back, $Z'_K = \mathrm{enc}(\mathrm{scene}|_K)$ and $Z'_O = 0$; the removed slots stay in the model's input, marked by a learned token, and both the squared error and the regularizer use K only. *Move:* the object's group centers are translated by $\delta = u \odot e\, s$, with $u \sim \mathcal{U}(-1,1)^3$, e the object's extent and scale $s = 0.5$ in training (a sweep of s gives the psychometric curves), and the full scene is re-encoded; Gaussians are translated, not rotated. *Add:* the object is removed from its own scene; the input combines the re-encoded scene without it and the object encoded on its own, and the target is the original full-scene encoding. For kept tokens with true change $\Delta_i = Z'_i - Z_i$ and prediction $\hat{\Delta}_i$, relative L2 is

$$\frac{1}{|S|} \sum_{i \in S} \frac{\|\hat{\Delta}_i - \Delta_i\|_2}{\max(\|\Delta_i\|_2, 10^{-8})},$$

where S is the quarter of kept tokens with the largest $\|\Delta_i\|_2$, selected within each evaluation minibatch of transitions (32; 16 in the 4096-token space); minibatch values are averaged with weights $|S|$. The identity prediction $\hat{\Delta} = 0$ scores exactly 1. Direction cosine is the mean $\cos(\hat{\Delta}_i, \Delta_i)$ over the same S, and beats-identity is the fraction of S with $\|\hat{\Delta}_i - \Delta_i\| < \|\Delta_i\|$. Unless stated otherwise, reported values average three world-model seeds. For a removal, S is drawn from the kept tokens. For a move, the same metrics are computed with S drawn from all tokens (the moved object's included), from the other tokens and from the object's own tokens; the move sweep of Appendix G reports the all-token values.

B.1 Implementation details and design controls

Run identifiers. Table 5 maps run identifiers to model names; launch scripts and configuration files specify every further setting.

Table 5: Run identifiers and model names. Encoder tokens = groups × tokens per group. Suffixes combine with any identifier.

identifier	model
`sketchF`	base recipe: one token per group, 512 groups, SIGReg, no image teacher
`reconF` / `reconV2`	masked parameter reconstruction, bare / with SIGReg
`visregRF2048x2`	reference: routed geometry/appearance pairs, 2048 groups × 2, VISReg-inspired hybrid, image teachers on the appearance token
`visregRF2048x2rope`	rotary model: the reference with 3D RoPE in the predictor's cross-attention
`visregRF1024x2`	sibling: the reference recipe at 1024 groups (2048 tokens)
`visregRF4096x2`	8192-token model: 4096 groups × 2, k=16, 200k-Gaussian cap
`visregSingle2048`, `visregSingle4096`	one token per group at 2048 / 4096 groups (4096: k=16); the teacher-weight sweep varies `visregSingle2048`
`bothRF1024x2`	token pairs at 1024 groups with both tokens distilled, SIGReg
`factorRF256`, `factorRF` ("factorRF-1024" in the tables), `factorRF1024x2`	routed pairs with SIGReg and real-frame teachers at 512-, 1024- and 2048-token budgets
`factorRF128`, `dupRF256`, `chunkRF256`	the same family at 256 tokens; 512-token controls with two full-width tokenizers (duplicate) or one shared tokenizer split into two tokens (chunked)
`dualRF`, `dualRFfull`, `dualRFg1024`	one token per group with real-frame teachers: 512 tokens on 260 / 400 scenes; 1024 tokens
`ssl`	no image-teacher losses (`visregRF2048x2ssl` is the image-free pair)
`vx`	evaluation scenes excluded from pretraining (274 training / 64 validation scenes)
`corpus` / `droom` / `indfull`	inductive 1056-scan / 645-room / 1381-scan corpora (Table 6)
`km` / `cc` / `nomask`	k-means proxy / connected-components proxy / no object or part masks

Inputs and architecture. Scenes are subsampled to at most 100k Gaussians (200k for the k=16 models). The raw member vector is $[\mu - \mu_{\text{ctr}} \,|\, s \,|\, q \,|\, o \,|\, c_{dc} \,|\, c_{sh}] \in \mathbb{R}^{59}$: relative position, log-scale, quaternion, opacity logit, SH-DC and degree-three SH-rest, with dimensions $3/3/4/1/3/45$. The single-token PointNet is a per-member MLP $59 \to 128 \to 128 \to 768$ (GELU), max-pooled over members and layer-normalized; routed models use two such PointNets, on geometry (11: relative position, log-scale, quaternion, opacity) and appearance (48: SH-DC and SH-rest), each token with a learned type embedding. The encoder has width 768 and 12 layers; the predictor has width 384 and six layers. Position and local content enter separately: each PointNet sees member coordinates relative to its group center, the standardized center enters through a sinusoidal encoding, and the predictor queries masked groups at their supplied centers.

Masking and masked fractions. Spatial-block, object-aware and random scene masks have effective weights $0.5/0.375/0.125$; part masks apply only to object-level inputs. Object-aware masks come from ground-truth instance partitions without class labels, and both tokens of a group are masked together. The configured mask ratio is 0.60, but object-mask sampling has bounds 0.30–0.80, so realized masks vary around it. On the base recipe, object-aware masking beats random masking (0.1074 vs. 0.0965, original probe split, n=5 paired seeds, t=3.96, $p\approx$0.017), and spatial-block masking is indistinguishable from random.

Label-free proxy masks. Three controls on the reference recipe replace the instance partitions by k-means proxies (k=30, the corpus-mean instance count), connected-components proxies, or no object or part masks; with proxy masks, pretraining uses no annotation of any scene. The probe is flat across the four regimes (0.533–0.540), while the scene-mean margin orders annotated $>$ k-means $>$ connected components $>$ none (0.836, 0.785, 0.759, 0.699; three seeds each): the object-shaped mask prior is worth 0.14 of margin, and label-free proxies recover 0.06–0.09 of it. Zero-shot on 20 InteriorGS scenes, the reference keeps a margin of 0.600 ± 0.019 over the scene-mean null (three seeds), 72–77% of the in-domain margin across the reference and sibling encoders.

Anti-collapse without a momentum teacher. Collapse is diagnosed in encoder space (effective rank, per-dimension standard deviation, α-ReQ spectral decay, LiDAR), and collapsed runs are kept only as labeled controls. On the base recipe (original probe split), SIGReg alone beats the VISReg-inspired hybrid (0.1059 ± 0.0038 vs. 0.0934 ± 0.0034, n=5, paired t=4.06), an EMA teacher collapses (effective rank 44; probe 0.0794), and VICReg [12] reaches near-full rank ($764/768$) at the lowest probe (Table 8), so rank alone does not certify quality.

Unabridged text, the mask-source table and the backbone, loss, input-feature and positional-encoding ablations: Extended results, Section 1 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

C Evaluation protocols, statistics and compute

C.1 Splits and exposure

Recognition and prediction use different scene sets (Table 6). The probe is trained with class labels on 338 scenes and evaluated on the 62 official validation scenes (per-Gaussian voted mIoU); it standardizes its input features, without which even SceneSplat's pretrained features underfit at 0.03. CF and its margins use 20 fixed validation scenes of the 400-scene corpus and the largest labeled instance per scene; some base-recipe ablations use a 10-scene version and say so. Exposure means that a scene's splats, and for distilled recipes its photo-lifted teacher targets, enter pretraining without class labels; instance boundaries enter only through object masks, which label-free proxy masks remove. The 20 CF scenes receive no pretraining gradient on the controlled, val-excluded and 1056-scan corpora; the 645-room and 1381-scan corpora contain 17 of them and the 1513-scan corpus all 20, so for the 645-room and 1381-scan corpora *inductive* qualifies the probe only. InteriorGS experiments use the first 200 of its 1,000 scenes, by identifier.

Probe exposure. With the 62 evaluation scenes excluded from pretraining (274 training scenes instead of 320), the reference recipe probes 0.2855 ± 0.0096 (n=3) against 0.5291 ± 0.0073 (n=5), retaining 54%, at a similar margin (0.8372 ± 0.0058 vs. 0.8295 ± 0.0128, Welch p=0.29); the 8192-

Table 6: Pretraining corpora and exposure of the two evaluation sets. Probe rooms: the 62 official validation scenes. CF scenes: the 20 scenes of the prediction protocol; counts give how many are in the pretraining split. The 80 untouched official-validation rooms (Table 9) are outside the 400-scene corpus and its probe splits and inside the 1056–1513-scan corpora.

corpus	pretraining split	probe rooms	CF scenes	models
controlled, 400 ScanNet-MCMC scenes	320 train / 80 validation	inside (transductive): 46 in train, 16 in validation	0/20 (validation split)	base, reference, rotary, frontier sweep
controlled, evaluation scenes excluded	274 train / 64 validation	excluded (inductive)	0/20	val-excluded controls
1056 scans	evaluation rooms and their rescans excluded	excluded (inductive)	0/20	corpus-scaling rows
645 rooms, one scan each	evaluation rooms excluded	excluded (inductive)	17/20	corpus-scaling rows
1381 scans of 645 rooms	every capture except the evaluation rooms	excluded (inductive)	17/20	inductive reference and rotary rows
1513 scans	includes the evaluation rooms	inside (transductive)	20/20	rotary contrast; target-removal control

token model retains 57% (0.3127 ± 0.0081), $+0.027$ over the reference control (Welch p=0.02, n=3 each). Distilled recipes keep 54–69% of their transductive probe with raw CF within 0.01, while the image-free base recipe does not drop (0.1148 ± 0.0055 vs. 0.1059 ± 0.0018; Figure 5). More inductive data raises the distilled probe (the reference recipe rises $0.286 \rightarrow 0.384 \rightarrow 0.433 \rightarrow 0.448$ from 274 to 1381 scans, n=3 each) but not the image-free one (the 2048-token pair recipe without image teachers stays at 0.130–0.138 from 274 to 645 rooms). The smaller corpus predicts only 0.01–0.03 of the reference's gap (shrinking a render-target recipe from 319 to 208 scenes costs 0.031 mIoU), and a target-removal control locates the rest: keeping the evaluation rooms' splats but removing their photo-lifted targets gives 0.443 ± 0.009 on 1513 scans, level with the inductive 0.448 (matched steps, three seeds each; 0.672 with the targets).

C.2 The probe gap against a measured ceiling

The oracle mIoU of our token partition (majority ground-truth label per token, official-val protocol) is 0.9262 at 512 tokens (0.9196 / 0.9290 / 0.9275 at 256 / 1024 / 2048), and token purity is 0.968–0.973: this oracle-overall-accuracy bound [39] is flat where the probe moves most, so granularity gains come from the representation, not from purer segments. Our models sit at about 11–61% of the matching-granularity ceiling, and the reference reaches 81% (0.529) of the photo-lifted teacher features' probe (0.654, Table 4) transductively and 44% (0.286) inductively (Appendix D.4).

C.3 Statistics

Units and intervals. By default, $\pm$ is a sample ($n-1$) standard deviation over encoder (pretraining) seeds, with n beside the value. Frozen-feature predictor comparisons use three predictor seeds on one encoder, world-model results three world-model seeds, and the fixed-frame decomposition and paired per-scene tests use scenes; few-shot accuracy averages 100 fixed episodes per checkpoint before the seed-level standard deviation. Retrieval reports paired 90% intervals over scenes, move thresholds scene-level bootstrap 95% intervals, and edit search 95% Wilson and scene-bootstrap intervals.

Detectable differences and multiplicity. With the reference's spread, the smallest difference detectable between two n=5 arms (two-sided α=0.05, 80% power, Welch) is 0.015 probe mIoU and 0.026 margin (0.018 / 0.036 between n=3 arms). The frontier sweep holds 32 arms selected on the same 62 scenes (496 pairwise contrasts), so we read it as a map and report paired tests descriptively, without family-wise correction. Four seed-level contrasts the text relies on survive a Holm correc-

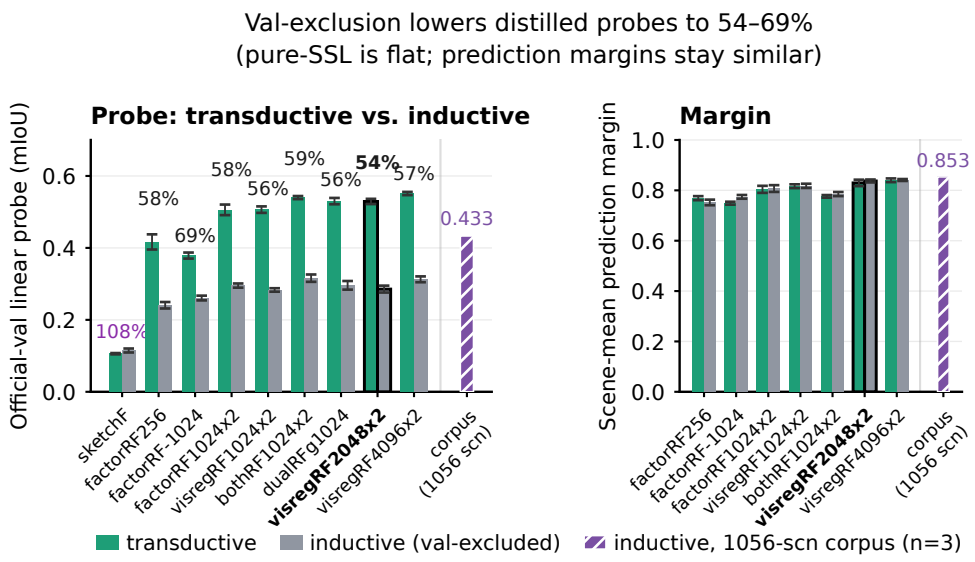


Figure 5: **Excluding the evaluation scenes from pretraining.** Left: official-val probe after transductive and inductive (val-excluded) pretraining, per arm (five encoder seeds for the transductive base, reference, sibling and 8192-token values, three otherwise; retained fraction above each pair; hatched: the reference recipe on the inductive 1056-scan corpus). Right: scene-mean prediction margin of the same arms. Exclusion lowers every distilled probe to 54–69% of its transductive value; the base recipe and the margins stay similar.

tion over the four: corpus scaling on the inductive split (Welch t=24.6, p=3×10^{-4}), the k-means masking margin cost (t=−5.2, p=0.001), the 8192-token probe gain (t=5.8, p=8×10^{-4}) and the reference's probe gain over its sibling (t=4.4, p=0.003); the corresponding margin differences (p=0.16 and p=0.09) are within seed noise. Within-axis tests such as the VISReg margin gain are paired per scene over the 20 CF scenes on seed-averaged margins, so they estimate variation across scenes, not training runs.

Selection on the reporting scenes. Recipes were selected on the 62 official-val scenes we report. Over 2000 random half-splits of the pooled per-scene confusion matrices of six frontier arms, selecting on one half inflates the picked arm's score there by 0.010 mIoU, below the 0.015 detectable difference (0.023 at the 90th percentile). The reference ranks fifth of the six and is selected in 0.0% of half-splits, so the probe ordering did not choose it; the arms that outrank it are reported.

C.4 Compute and model size

Table 7 lists parameter counts and pretraining wall-clock; the card varied by arm (24–48 GB), so wall-clock is not a matched efficiency comparison. Probe and CF evaluation add 10–16 minutes per arm, each world-model training takes 13–44 minutes, and teacher features cost 36 s per scene; Gaussian fitting (3DGS-MCMC [27]) is not included. On one L40S, a shared-encoder step takes 0.40 s and 10.8 GiB, against 0.305 s and 6.7 GiB for an EMA teacher with detached targets, which adds 85.1 million parameters and changes the objective. Unabridged text, the exposure ladder and matched-split controls: Extended results, Section 2; every arm's wall-clock and the recipe-cost diagnostic: Section 3 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

Table 7: Compute and model size. Parameter counts are instantiated from the reference model's saved configuration; the context and target encoders are the same network. Wall-clock is the elapsed time of the seed-42 pretraining run, or the range over the seeds stated in the row, on the card the scheduler assigned. Three arms; Extended results, Section 3 lists every arm.

module (reference model, visregRF2048x2; identical module set for visregRF1024x2)	params	
factorized tokenizer (geometry + appearance PointNets)	0.24 M	
ViT-B encoder (shared context/target)	85.06 M	
cross-attention predictor (6 layers, 384-d)	14.80 M	
DINOv2 + SigLIP2 distillation heads	2.36 M	
total	**102.46 M**	
edit world model (remove or move head, 4 layers, 768-d)	28.9 M each	
arm (seed 42)	**GPU**	**wall-clock**
sketchF (base recipe, 512 tokens)	L40S 48 GB	0 h 36
visregRF2048x2 (reference, 2048×2 tokens, batch 4)	48 GB class (L40S/A40/A6000)	3 h 25–9 h 02 (five seeds)
visregRF4096x2 (8192 tokens, k=16, batch 4)	48 GB class	8 h 16–15 h 02 (five seeds)

D Recognition and prediction

D.1 Full results

Seed spread for Table 4. Transductive probe, CF and margin, with sample s.d. over encoder seeds: reference (n=5) 0.5291 ± 0.0073, 0.8860 ± 0.0112 and 0.830 ± 0.013; 2048-token routed pair (n=5) 0.5068 ± 0.0086, 0.8917 ± 0.0050 and 0.8164 ± 0.0075; both tokens distilled (n=3) 0.5407 ± 0.0055, 0.8905 ± 0.0060 and 0.7761 ± 0.0052; 8192 tokens (n=5) 0.5512 ± 0.0045, 0.8834 ± 0.0074 and 0.840 ± 0.007; single token at 4096 groups (n=3) probe 0.565 ± 0.020 and margin 0.789 ± 0.008. Inductive probes (n=3 each): base recipe 0.1148 ± 0.0055, 2048-token routed pair 0.2831 ± 0.0048, both tokens distilled 0.3159 ± 0.0102, reference 0.2855 ± 0.0096, 8192 tokens 0.3127 ± 0.0081. On the 1056-scan corpus (n=3 each), the routed pair reaches 0.4104 ± 0.0088, 0.8997 ± 0.0067 and 0.8474 ± 0.0080, and the reference recipe 0.4331 ± 0.0040, 0.8984 ± 0.0033 and 0.8534 ± 0.0100. The 1381- and 1513-scan rows and the rotary model are in Table 13; one-scorer margins for the controlled-corpus JEPA rows are in Tables 11 and 12.

Non-JEPA rows and base-recipe CF. The non-JEPA rows of Table 4 are single runs under the same probe: DINOv2+SigLIP2 features lifted onto the same groups with no JEPA, from the original photographs (0.654) or from splat renders (0.481); SceneSplat's public vision–language checkpoint with all 338 probe-training scenes (0.651; 0.607 on the original probe split, Table 8); and the 2048-token partition oracle (Appendix C.2). The base recipe's CF of 0.848 (Tables 2 and 4) comes from one evaluation on 20 scenes, available as an aggregate only; the same recipe pretrained and scored on InteriorGS reaches 0.813 (15 scenes).

Foreground mIoU (wall and floor excluded), per-class IoU against the lifted features and a raw-parameter probe: Extended results, Section 4 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

D.2 Unseen rooms

Of the 142 rooms in ScanNet's official validation list, 80 are outside the 400-scene corpus and its probe splits, so the controlled-corpus models never saw them in pretraining, probe fitting or selection (the 1056–1513-scan corpora contain them). Each checkpoint is probed on each room's first scan and, with the same probe seed, on the 62 official rooms (Table 9). Paired by seed, the rotary model exceeds the reference on the untouched rooms by $+0.021\pm0.003$ (all five seeds, $+0.016$ to $+0.024$). The lower scores of both distilled recipes on these rooms combine room difficulty, exposure and selection history; the comparison was not pre-registered and is descriptive.

Table 8: Base-recipe family (512 tokens) and controls: voted mIoU on ScanNet-20 and 5-way 10-shot ModelNet40 accuracy (sample s.d. over seeds). "Official" is the official ScanNet-v2 validation protocol; these scenes also informed recipe selection. FA-yaw: frame-averaging wrapper (Extended results, Section 8, `https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`). visregF: the VISReg-inspired hybrid of Appendix B in place of SIGReg. SceneSplat: its public, language-distilled checkpoint under our probe, an upper reference. †Earlier few-shot evaluation (Appendix D.3).

Arm	mIoU (probe split)	mIoU (official)	few-shot 5w10s
sketchF (ours: SIGReg, no EMA)	0.1059 ± 0.0038 (n=5)	0.1059 ± 0.0018 (n=5)	$77.1 \pm 6.2\%$ (n=5)
+ FA-yaw wrapper	0.1222	—	—
+ FA-yaw × multi-block stack	—	0.1480	—
tokauxNCE + stack (pure-SSL config)	—	0.1479 ± 0.0067 (n=3)	—
dinoauxNCE (DINOv2 distillation aux, full weight)	—	0.2777 ± 0.0116 (n=3)	—
dinoW05 (DINOv2 aux, weight 0.5; single channel)	—	0.2479 ± 0.0015 (n=3)	—
+ FA-yaw	—	0.2779 (n=1)	—
dualW05 (DINO+SigLIP2 head)	—	$\mathbf{0.2566 \pm 0.0040}$ (n=3)	—
siglipW05 (SigLIP2 head, weight 0.5)	—	0.2146 (n=1)	—
visregF (hinge+SIGReg hybrid)	0.0934 ± 0.0034 (n=5)	—	—
hingeF (VICReg)	0.0324 ± 0.0008	0.0322	40.8% †
reconF (masked parameter reconstruction)	0.0979 ± 0.0026	0.1100 (n=1)	35.5% (n=1)
emaF (EMA teacher, base-recipe settings)	0.0794 ± 0.0014	0.0854 (n=1)	—
SceneSplat features (VL-supervised)	0.607	0.651	—

Table 9: Probe on 80 untouched rooms vs. the 62 official rooms, same probe seed (mean ± sample s.d. over encoder seeds). The base-recipe row is one checkpoint re-probed with this table's harness; its official-room value differs from the five-seed mean of Table 2.

model	seeds	untouched 80	official 62	paired difference
reference	5	0.345 ± 0.003	0.529 ± 0.006	-0.184 ± 0.008
rotary	5	0.366 ± 0.005	0.535 ± 0.014	-0.169 ± 0.015
base recipe (no image teacher)	1	0.129	0.120	$+0.009$

D.3 Matched controls

Few-shot protocol and provenance. Each episode draws 5 classes and, per class, 10 support and 20 query objects without replacement from the 22 classes with at least 30 test objects among the 2,467 ModelNet40 test-split splats; a query takes the label of the nearest class-mean prototype of concatenated max- and mean-pooled tokens. One fixed manifest of 100 episodes (seed 0) serves every checkpoint; values are episode means per checkpoint, then mean ± s.d. over model seeds (latent $77.1 \pm 6.2\%$ over five seeds; SIGReg-regularized reconstruction $91.1 \pm 0.4\%$ over three). These three regularized reconstruction checkpoints (`reconV2`) pass collapse checks and probe 0.137–0.151 (Table 2). The bare checkpoint (one seed) predates both faithfulness fixes of the regularized recipe (canonical member ordering, equal per-attribute loss weights), so the reconstruction rows differ as recipes. The hinge (VICReg) few-shot entry of Table 8 is from an earlier evaluation.

Reconstruction-path counterfactual. Reconstruction arms never train the predictor output that CF scores, so we also score them in their own output space: the trained reconstruction head decodes the masked groups' predictor output into raw 59-d parameters, which are spliced into the scene and re-encoded by the same frozen encoder, and the masked groups' embeddings are compared between the spliced and true scenes; the nulls (scene-mean parameters, a random other group) follow the same path. On 15 validation scenes the three regularized checkpoints score 0.441 ± 0.020 against 0.375 ± 0.029 and 0.363 ± 0.006 ($+0.066$ and $+0.078$, every seed positive); the bare checkpoint scores 0.994 against 0.993 and 0.993, consistent with its effective rank of 178–205.

Fresh predictors on the frozen 2×2 encoders. A new predictor is trained for four frozen encoders of Table 3 (reconstruction + SIGReg, distillation only, JEPA only, JEPA + distillation) on a $208/52$-scene split with the matched-predictor protocol of Table 12 (standardized targets, final step; three predictor seeds). In the primary *visible-input* protocol the context encoder sees only the visible groups; the predictor still receives the masked-group centers and grouping, so the protocol

removes the attention path from masked groups to context tokens, not every geometric cue. In the *cached-feature* protocol the context features have attended to the masked groups, and mask sampling, position normalization and context standardization also differ. The JEPA-only encoder keeps the largest visible-input CF and random-token and scene-mean margins (Table 10), an order that raw-target and best-validation diagnostics preserve; the encoders differ as recipes, so the comparison is recipe-level.

Table 10: Fresh matched predictors on four frozen encoders of Table 3, with context from visible groups only (primary) or from cached full-scene features (not run for the reconstruction encoder). Mean ± s.d. over three predictor seeds on one frozen encoder per recipe; margins against each null. The protocols also differ in mask sampling, position normalization and context standardization (text).

	visible-input context				cached full-scene context	
encoder	CF	vs. random	vs. scene mean	vs. shuffled	CF	vs. random
recon. + SIGReg	0.562 ± 0.007	0.542 ± 0.008	0.486 ± 0.007	-0.002 ± 0.005	—	—
distill. only	0.675 ± 0.015	0.503 ± 0.016	0.380 ± 0.015	0.016 ± 0.004	0.818 ± 0.002	0.571 ± 0.001
JEPA only	0.877 ± 0.002	0.855 ± 0.002	0.780 ± 0.002	0.010 ± 0.002	0.907 ± 0.002	0.883 ± 0.002
JEPA + distill.	0.694 ± 0.004	0.599 ± 0.005	0.504 ± 0.004	0.009 ± 0.001	0.803 ± 0.003	0.682 ± 0.003

D.4 THE DISTILLATION FRONTIER

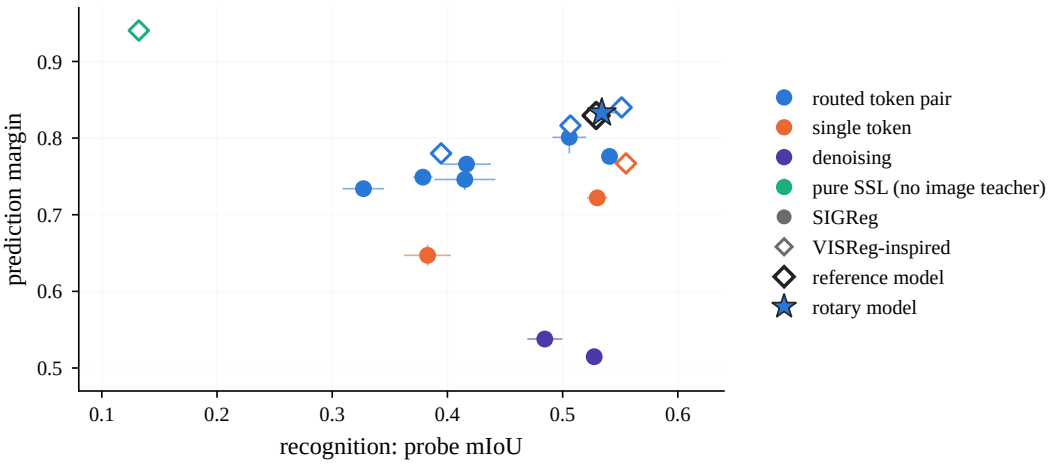


Figure 6: **Recognition and prediction trade off; at every matched token count, routed pairs have higher margins than single tokens.** Official-val probe against scene-mean prediction margin, one point per model (filled: SIGReg; open: VISReg-inspired). At matched token count, routed pairs (blue) have higher margins than single tokens (orange), at 512 tokens with a higher probe as well; denoising models (violet) have the lowest margins, and the image-free pair (aqua; "pure SSL" in the legend) the highest margin and lowest probe. The rotary model (star) sits with the reference (black outline); the rightmost blue diamond is the 8192-token pair. Margins compare models at one token count (Appendix D.5). Sample s.d.; n=1–5; transductive.

The reference combines real-frame teachers, routed token pairs, the VISReg-inspired hybrid and 4096 tokens; below are the matched comparisons behind each choice (*margin*: CF minus the scene-mean null). The CF $\geq$ 0.90 bar of Section 3.2 was fixed as the adoption criterion for distilled recipes before the first image-teacher arm was trained.

The teacher-weight trade-off. Raising the image-teacher weight α of the single-token 2048-group recipe from 0 to 2.5 raises the probe from 0.127 to 0.557 and, under one scorer (Table 11; n=2–3 per setting), lowers the scene-mean margin from 0.917 ± 0.008 to 0.716 ± 0.004 and the random-token margin from 0.923 ± 0.010 to 0.823 ± 0.001, monotonically across the six settings

(main-sweep values in Figure 3: $0.915 \rightarrow 0.717$ and $0.920 \rightarrow 0.824$). Re-scoring in a fixed, per-dimension standardized space moves no margin by more than 0.02, so anisotropy does not drive the fall. A shuffled-teacher control (cached teacher rows permuted across groups within each scene; α=1 and 2.5, one seed each) lifts the probe only from 0.127 to 0.20 (18% of the content arm's gain) and keeps the random-token (0.90–0.92) and nearest-visible (0.59–0.60) margins at their teacher-free level, while its scene-mean margin falls to 0.82, about 60% of the content arm's fall: the probe gain and the losses under content-sensitive nulls come from the teacher's content, and part of the scene-mean margin's fall from the extra regression term itself.

Masked denoising and image-free self-distillation. Replacing the learned mask queries with noised, t-conditioned target embeddings (x0-prediction [2], following N-JEPA [1]) moves further along the trade-off (256 groups, n=3: probe 0.4845 ± 0.0152, margin 0.5378 ± 0.0093). A Sonata-style [31] self-distillation term, in which the visible-group context pass matches an EMA teacher's Sinkhorn-balanced prototype assignment of each visible token, raises the image-free pair recipe on the inductive 1381-scan corpus from 0.136 ± 0.006 to 0.188 ± 0.004 (n=3 each) at CF 0.922 and margin 0.930 (Table 4).

Real-frame teachers. A dual DINOv2+SigLIP2 head distilling features lifted from rendered views raises the base recipe to probe 0.2566 ± 0.0040 at CF 0.897 (n=3). The lifted features bound the student: with no JEPA, lifted DINOv2 features probe 0.458 from renders and 0.607 from real frames at their poses on the same 260 scenes.

Routed token pairs. At a matched 512-token budget (256 groups, n=3 per arm), three ways of producing a routed pair all sit above the single token (probe 0.3828 ± 0.0203, CF 0.8877): separate geometry and appearance tokenizers (factorized; 0.4166 ± 0.0211, 0.8968 ± 0.0044), two full-width tokenizers (duplicate; 0.4128 ± 0.0029, 0.8980 ± 0.0014) and one tokenizer whose output is split into two tokens at about $1\times$ capacity (chunked; 0.4151 ± 0.0265, 0.9034 ± 0.0034). Two tokens per group with routed distillation, not tokenizer capacity, account for the gain, and the scene-mean margin separates the arms more clearly (0.652 single token, one run, against 0.746–0.766; Table 12). Routing both teachers into the geometry token (n=3) gives it a probe of 0.4114 ± 0.0122, close to the appearance token's 0.420, while the freed appearance token falls to 0.1974 ± 0.0091: distillation sets the probe, and the undistilled token carries the higher margin.

Finer granularity. In the main sweep, at 1024 groups under SIGReg (n=3 per arm), a second token distilled like the first raises the margin from 0.722 ± 0.005 to 0.7761 ± 0.0052, and an undistilled second token raises it to 0.804 ± 0.017 while lowering the probe relative to the both-distilled pair from 0.5407 ± 0.0055 to 0.5058 ± 0.0147. At 2048 tokens, one token per group (the reference recipe with the factorization off, n=3) probes 0.5550 ± 0.0089, above every 2048-token pair and the reference (Welch $p \leq 0.02$ against the sibling and against the reference), and lies below them on CF and both global-null margins (CF 0.861 ± 0.012, margin 0.767 ± 0.014, random-token margin 0.847 ± 0.012; Welch $p \leq 0.041$ on all three against the sibling, n=5). The 4096-group single token (k=16, n=3) is again below the pairs at matched token or group count on the margin (0.789 vs. 0.830 for the reference and 0.840 for the 8192-token pair). Context-sensitive nulls do not favor pairs (mean-context margin 0.316 ± 0.026 for the 2048-group single token, 0.271 ± 0.048 for the sibling; Table 11).

Regularizer. At 1024 groups, the VISReg-inspired hybrid (not the published VISReg objective [13]) and our SIGReg give margins of 0.728 ± 0.009 and 0.722 ± 0.005 for single tokens and 0.816 ± 0.008 and 0.804 ± 0.017 for pairs (n=3 per cell; n=5 for the VISReg pair). Paired per scene over the three matched seeds, the VISReg gain is $+0.005$ on single tokens (t=1.5) and $+0.0136$ on pairs (t=3.59, df=19, $18/20$ scenes; not significant across seeds); the difference between the two gains ($+0.008$, t=1.7) is not established, so we claim no regularizer-by-pairing interaction.

Unabridged text, predicted-token probing along the dial, per-member margins of routed pairs, distillation coverage and teacher variants: Extended results, Section 5 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

D.5 MARGIN DEFINITION AND NULLS

Margins and nulls. CF is the cosine between the predictor's output at a masked group and that group's full-scene target, scored on the largest labeled instance of each of 20 validation scenes; each margin subtracts a no-prediction null computed on the same masked sets. The *scene-mean* null scores the mean context embedding against each target. The *identity* null is 1.0000 for every arm with the shared, momentum-free target encoder (its parameter-free target normalization is idempotent), so it is uninformative. The *random-token* null scores the prediction against the true embedding of a random other token of the same scene (eight draws), a specificity test. The *shuffled-context* null decouples context content from position, and the *mean-context* null replaces every context token's content by the scene-mean context embedding while keeping positions; both rerun the predictor. The *nearest-visible copy* scores the context-pass embedding of the nearest visible token (type-matched for pairs). Matched against every labeled instance of its scene (21–31 candidates), each predicted instance picks the true one in $360/360$ episodes across eighteen checkpoints: predictions are instance-specific.

Margins compare models at one token count. The scene-mean null falls with token count (0.173 at 256 tokens, three seeds, to 0.056 for the 4096-token reference, five seeds), so raw CF and margin rank granularities differently: the coarsest factorized pair has the highest raw CF of its family (0.908 ± 0.003) and the lowest margin (0.734 ± 0.009; Table 12). Under the random-token null, pair models span only 0.858–0.883. Replacing context content by its scene mean costs the reference 0.27 ± 0.05 of CF ($0.883 \rightarrow 0.614 \pm 0.038$, n=5), so the predictor reads context content; the shuffled-context margin is at most 0.011 for every JEPA row of Table 12, so the reference predictor is insensitive to how content is bound to position, which the rotary predictor changes (Appendix E).

Two protocols and a local baseline. The all-instance protocol re-scores the same scenes with every labeled instance of at least three groups and at most half the scene (657 per encoder seed at 4096 tokens); absolute CF falls with granularity under it, so only within-family contrasts are comparable, and at matched group count routed pairs keep more CF and scene-mean margin than single tokens under both protocols. The nearest-visible copy is a strong local baseline in every image-supervised space (cosine 0.60–0.65 against the truth over all instances): the reference beats it by 0.43 on the largest instance per scene and loses by 0.04 over all instances (Table 13), winning above about 40 groups (78% of instances at 41–80 groups, 98% above 80) and losing below 20 (0–6%; Figure 7). The margin is therefore a large-object measure, and the copy is the null a predictor must clear on small objects.

Table 11: **Prediction margins under context-sensitive nulls, one scorer.** Mean ± sample s.d. over n encoder seeds on the 20 CF scenes (Table 6), which the evaluation-scene-excluded models never trained on; largest labeled instance per scene. Probe: official-val mIoU of the teacher-weight settings plotted in Figure 3 (α=1 is visregSingle2048); other arms' probes are in Table 4 and Appendices D.1 and D.4. Extended results, Section 6 gives the all-instance protocol for every arm and the val-excluded and mask-source rows.

arm	n	probe	CF	vs. scene mean	vs. rand. token	vs. mean ctx.	vs. nearest vis.
factorRF128 (256 tok)	3	—	0.908 ± 0.003	0.734 ± 0.009	0.858 ± 0.007	0.141 ± 0.007	0.345 ± 0.005
factorRF256 (512)	3	—	0.896 ± 0.005	0.766 ± 0.010	0.871 ± 0.001	0.153 ± 0.008	0.354 ± 0.006
chunkRF256 (512)	3	—	0.902 ± 0.002	0.746 ± 0.014	0.868 ± 0.007	0.148 ± 0.011	0.363 ± 0.005
dupRF256 (512)	3	—	0.896 ± 0.001	0.766 ± 0.007	0.870 ± 0.006	0.150 ± 0.005	0.349 ± 0.008
dualRFfull (512, single)	1	—	0.863	0.652	0.811	0.164	0.318
factorRF1024x2 (2048)	3	—	0.881 ± 0.016	0.801 ± 0.021	0.871 ± 0.017	0.244 ± 0.056	0.391 ± 0.020
visregRF1024x2 (2048, sibling)	5	—	0.891 ± 0.007	0.815 ± 0.008	0.883 ± 0.007	0.271 ± 0.048	0.393 ± 0.013
visregRF2048x2 (4096, reference)	5	—	0.883 ± 0.014	0.827 ± 0.015	0.877 ± 0.014	0.269 ± 0.048	0.431 ± 0.010
bothRF1024x2 (2048)	3	—	0.891 ± 0.005	0.776 ± 0.005	0.870 ± 0.006	0.254 ± 0.060	0.403 ± 0.006
dualRFg1024 (1024, single)	3	—	0.869 ± 0.006	0.722 ± 0.004	0.838 ± 0.004	0.207 ± 0.044	0.382 ± 0.003
visregSingle2048 (2048, single)	3	0.555 ± 0.009	0.855 ± 0.007	0.761 ± 0.010	0.841 ± 0.008	0.316 ± 0.026	0.423 ± 0.009
visregSingle4096 (4096, single)	3	—	0.865 ± 0.010	0.784 ± 0.010	0.853 ± 0.010	0.386 ± 0.029	0.472 ± 0.015
visregRF4096x2 (8192)	5	—	0.883 ± 0.008	0.840 ± 0.007	0.879 ± 0.008	0.284 ± 0.042	0.475 ± 0.009
single token, α=0	3	0.127 ± 0.002	0.927 ± 0.009	0.917 ± 0.008	0.923 ± 0.010	0.360 ± 0.074	0.609 ± 0.001
single token, α=0.06	2	0.373 ± 0.000	0.914 ± 0.015	0.875 ± 0.017	0.907 ± 0.015	0.272 ± 0.025	0.587 ± 0.016
single token, α=0.15	3	0.474 ± 0.005	0.905 ± 0.004	0.845 ± 0.004	0.896 ± 0.004	0.252 ± 0.035	0.552 ± 0.000
single token, α=0.4	2	0.521 ± 0.002	0.880 ± 0.022	0.795 ± 0.021	0.866 ± 0.022	0.312 ± 0.036	0.496 ± 0.024
single token, α=2.5	2	0.557 ± 0.001	0.846 ± 0.001	0.716 ± 0.004	0.823 ± 0.001	0.310 ± 0.006	0.346 ± 0.013
image-free pair, 4096 tokens	3	—	0.947 ± 0.005	0.942 ± 0.004	0.945 ± 0.005	0.394 ± 0.075	0.639 ± 0.008

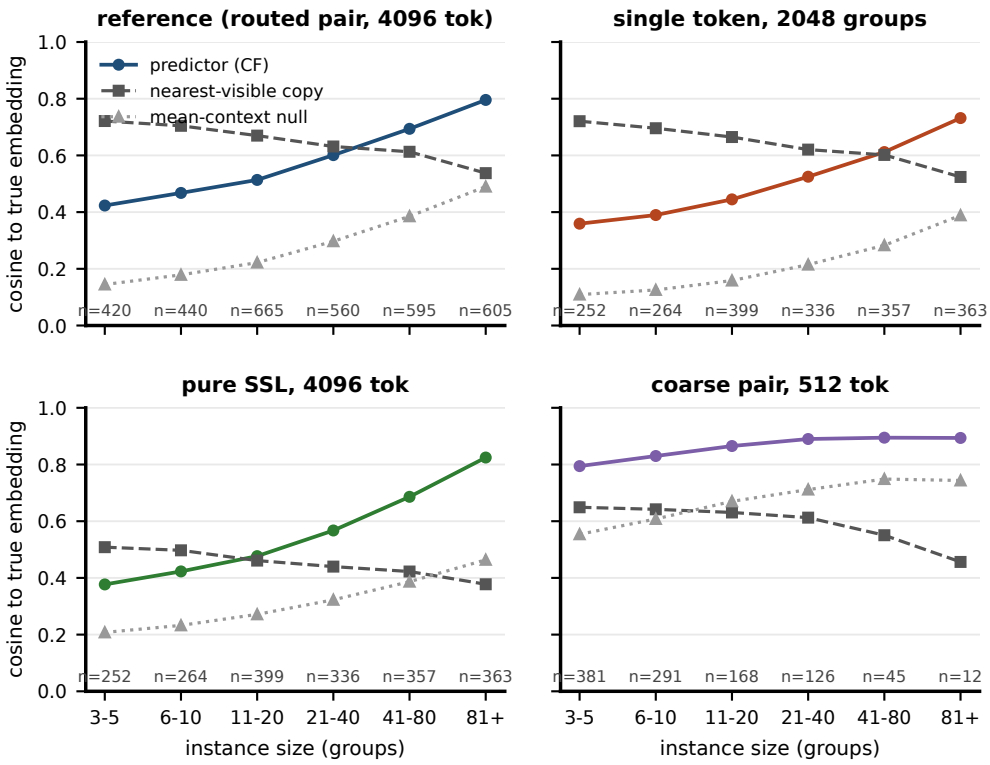


Figure 7: **Where prediction beats a local copy.** Per instance-size bin (groups): predictor CF, a copy of the nearest visible token's context embedding, and the mean-context null, over every labeled instance of the 20 CF scenes (seeds pooled; n: instances per bin). The reference and single-token model overtake the copy at 41–80 groups, the image-free model at 11–20; the coarse pair is above it at every size.

A shared null battery. Other representations are mean-pooled over the reference's 2048 groups and scored with a predictor matching the reference's architecture, masker, loss and 24k-step budget (320 training scenes, three predictor seeds; Table 12). Raw SceneSplat vision–language features have the smallest margins because two arbitrary groups already sit at cosine 0.93; standardized, they stay below our frozen reference encoder under matched predictors, whose margins exceed the end-to-end reference's 0.830, so the protocol does not disadvantage frozen encoders. The lifted teachers out-probe our encoder and retrieve the right class more often (Appendix F; Extended results, Section 9), yet reach far smaller margins: recognition and prediction rank these representations in opposite order. These margins measure completion of contextualized features (visible-input control: Table 10). SceneSplat's vision–language checkpoint is trained on about 7K scenes that include ScanNet training scenes; its self-supervised checkpoint runs with derived normals, and we draw no conclusion about that method from it.

Unabridged text, the all-instance protocol for every arm, margins by object class and size, the local copy along the teacher-weight dial and a relative-position-bias control: Extended results, Section 6 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

E Spatial Prediction

E.1 Rotary Position Encoding on Three Corpora

The rotary model differs from the reference only in its predictor: a 3D rotary position embedding [25], with mixed-axis frequencies generalized to xyz, acts on the queries and keys of the

Table 12: **Prediction margins under the null battery.** Mean ± sample s.d. over encoder seeds, three unless marked; 20 scenes, largest labeled instance per scene. Raw CF is the battery's own pass and can sit below the CF of Table 4; every margin uses it. JEPA rows share one scorer pass (factorRF-1024: a separate pass), which agrees with the pass behind Table 4 within 0.006; Table 11 adds their mean-context and nearest-visible margins. Frozen-feature rows (below the rule) use the matched-predictor harness, with ± over three predictor seeds on one frozen encoder; their context tokens have attended to the masked groups, so those margins measure completion of contextualized features, not inference from visible groups alone (the JEPA rows encode visible groups only).

			margin vs.		
arm or features	tokens	raw CF	scene mean	rand. token	shuf. ctx.
factorRF128	256	0.908 ± 0.003	0.734 ± 0.009	0.858 ± 0.007	0.002 ± 0.000
factorRF256	512	0.896 ± 0.005	0.766 ± 0.010	0.871 ± 0.001	0.004 ± 0.001
chunkRF256	512	0.902 ± 0.002	0.746 ± 0.014	0.868 ± 0.007	0.003 ± 0.001
dupRF256	512	0.896 ± 0.001	0.766 ± 0.007	0.870 ± 0.006	0.004 ± 0.000
factorRF-1024	1024	0.892 ± 0.001	0.749 ± 0.006	0.865 ± 0.001	0.004 ± 0.001
dualRFfull (n=1)	512	0.863	0.652	0.811	0.003
factorRF1024x2	2048	0.881 ± 0.016	0.801 ± 0.021	0.871 ± 0.017	0.007 ± 0.001
visregRF1024x2 (sibling, n=5)	2048	0.891 ± 0.007	0.815 ± 0.008	0.883 ± 0.007	0.006 ± 0.001
visregRF2048x2 (reference, n=5)	4096	0.883 ± 0.014	0.827 ± 0.015	0.877 ± 0.014	0.009 ± 0.001
bothRF1024x2	2048	0.891 ± 0.005	0.776 ± 0.005	0.870 ± 0.006	0.006 ± 0.000
dualRFg1024	1024	0.869 ± 0.006	0.722 ± 0.004	0.838 ± 0.004	0.008 ± 0.001
visregSingle2048 (single, 2048 groups)	2048	0.855 ± 0.007	0.761 ± 0.010	0.841 ± 0.008	0.008 ± 0.002
visregSingle4096 (single, 4096 groups)	4096	0.865 ± 0.010	0.784 ± 0.010	0.853 ± 0.010	0.009 ± 0.002
visregRF4096x2 (n=5)	8192	0.883 ± 0.008	0.840 ± 0.007	0.879 ± 0.008	0.011 ± 0.001
SceneSplat frozen (VL PTv3 [62])	2048	0.983 ± 0.000	0.021 ± 0.000	0.033 ± 0.000	0.007 ± 0.000
SceneSplat frozen, z-scored	2048	0.816 ± 0.002	0.587 ± 0.002	0.545 ± 0.001	0.167 ± 0.004
SceneSplat self-supervised (label-free; derived normals), z-scored	2048	0.835 ± 0.003	0.692 ± 0.003	0.597 ± 0.002	0.259 ± 0.003
lifted DINOv2, no JEPA (0.607-probe features), z-scored	2048	0.715 ± 0.002	0.560 ± 0.002	0.545 ± 0.003	0.191 ± 0.000
lifted SigLIP2, no JEPA, z-scored	2048	0.674 ± 0.002	0.369 ± 0.002	0.413 ± 0.003	0.136 ± 0.005
reference frozen, same harness	2048	0.945 ± 0.001	0.926 ± 0.001	0.932 ± 0.001	0.021 ± 0.001

predictor's cross-attention (fixed frequencies, wavelengths 0.25–16 normalized units). On the 400-, 1381- and 1513-scan corpora it raises the shuffled-context margin from 0.009–0.015 to 0.618–0.634 with similar probe and CF (on 400 scans the probe difference, paired over the five seeds, gives t=0.9, p=0.4) and similar or higher scene-mean margins (Table 13). On 400 scans the mean-context margin rises from 0.269 ± 0.048 (Table 11) to 0.752 ± 0.029 (n=5 each; the rotary value is from the per-run null pass). The pre-registered criterion (shuffled-context margin above 0.10; probe, CF and scene-mean margin no more than one reference standard deviation below the reference) is met by all nine pre-registered encoders (seeds 42, 1 and 2 on each corpus); the 400-scan means average five seeds. All eleven encoders pass collapse checks (effective rank 255–303, no dead dimensions).

E.2 Transplanted-object detection

Protocol. Into each official-val scene we transplant one labeled instance (8–60 groups) from another official-val scene by moving its group centers, which overwrites the resident groups nearest its new centroid. The transplant floats 1–1.8 m above the floor (*implausible*) or rests on the floor at a random free spot at its donor height (*plausible*, no geometric cue). The edited scene is re-encoded, as a deployment would, and every labeled instance is scored by *prediction error* (1− mean cosine between the predictor's output for the masked instance and its actual tokens) and by three predictor-free scores: *scene-mean distance* of its mean embedding to the context mean, *neighborhood dissimilarity* to the eight nearest groups of other instances, and *height* above the scene median. Table 14 reports AUROC with transplants as positives, pooled over scenes, and the per-scene top-1

Table 13: **The rotary contrast on three corpora.** Mean ± sample s.d. over three encoder seeds (400 scans: five). Probe: official ScanNet-20 val; *inductive* qualifies the probe column only. CF and margins: 20 CF scenes, largest instance per scene; CF of the 1381- and 1513-scan rows is from the null-battery pass. All-instance: every labeled instance of the same scenes (657 per encoder seed, pooled over the 20 scenes); CF − copy subtracts the nearest-visible-copy baseline. The 400-, 1381- and 1513-scan models have 0, 17 and 20 of the 20 CF scenes in their pretraining split (Table 6).

corpus	predictor	probe	CF	scene-mean margin	shuffled-ctx margin	all-inst. CF	CF − copy
400 scans (transductive)	reference	0.529 ± 0.007	0.886 ± 0.011	0.830 ± 0.013	0.009 ± 0.001	0.595 ± 0.018	-0.044 ± 0.016
	+ rotary (n=5)	0.534 ± 0.013	0.889 ± 0.005	0.833 ± 0.010	$\mathbf{0.632 \pm 0.033}$	$\mathbf{0.869 \pm 0.013}$	$\mathbf{+0.217 \pm 0.016}$
1381 scans (inductive)	reference recipe	0.448 ± 0.006	0.904 ± 0.004	0.857 ± 0.006	0.015 ± 0.001	0.658 ± 0.023	$+0.054 \pm 0.020$
	+ rotary	$\mathbf{0.460 \pm 0.015}$	0.905 ± 0.003	0.861 ± 0.005	$\mathbf{0.634 \pm 0.009}$	$\mathbf{0.891 \pm 0.006}$	$\mathbf{+0.266 \pm 0.005}$
1513 scans (transductive)	reference recipe	0.672 ± 0.006	0.903 ± 0.005	0.862 ± 0.005	0.015 ± 0.001	0.680 ± 0.036	$+0.079 \pm 0.037$
	+ rotary	0.670 ± 0.012	0.907 ± 0.002	$\mathbf{0.874 \pm 0.004}$	$\mathbf{0.618 \pm 0.028}$	$\mathbf{0.892 \pm 0.003}$	$\mathbf{+0.275 \pm 0.006}$

rate; restricting the negatives to transplant-sized instances moves no prediction-error AUROC by more than 0.04.

Result. The rotary predictor's error reaches AUROC 0.979 ± 0.005 floating and 0.884 ± 0.001 on the floor (n=3) and ranks the transplant first in 80% / 52% of scenes (seed 42), ahead of the static neighborhood score in the same latent. A height score alone reaches 0.97 on floating transplants, so the floor-placed case is the informative test. The reference, pure-SSL pair and single-token predictors do not separate transplants, for a structural reason: the reference's target encoder re-encodes the foreign object inside its new context and its predictor outputs a context-determined token, so both sides of the comparison move together. Matched predictors on frozen SceneSplat or lifted features reach 0.98–1.00 floating, but they score the transplant against tokens computed in its donor scene: scoring our reference encoder (seed 42) and its 1381-scan recipe that way gives 1.00, and scoring natives against displaced host-scene tokens gives 0.59–0.66, so that protocol flags any encoding mismatch.

Unabridged text, the rotary model's embedding map and a context-free re-encoding diagnostic of transplants: Extended results, Section 7; optional frame averaging of the frozen encoder: Section 8 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

F RETRIEVAL

Each masked group's predicted embedding queries a bank of real group tokens from *other* scenes (top-1 cosine), and the retrieved group's majority semantic class is compared with the masked group's own; because a correctly predicted class retrieves a different instance of that class, the readout is class agreement. The *oracle* queries with the encoder's true embeddings of the hidden groups; the nulls use the true embeddings permuted within the masked region, unrelated groups at least 1 m away, and the scene-mean embedding.

Population evidence. Over the 62 official validation scenes with three seeded object-level masks each (one 149-scene bank; Table 15), predicted tokens in the reference space agree with the hidden classes almost as often as the oracle and far more often than every null; retrieval cosine measures confidence, not correctness. The pattern holds for the reference recipe pretrained without the evaluation scenes (predicted 0.479 ± 0.119, oracle 0.483 ± 0.127, paired p=0.25) and for the 2048-token sibling (0.507 ± 0.116 vs. 0.511 ± 0.121, p=0.23).

Instance-level readout. Strong token-level retrieval does not imply equally strong retrieval of whole object instances. Voting one labeled instance per mask (majority over object-like tokens), the reference space matches the voted class at 0.204 ± 0.298 against the oracle's 0.242 ± 0.309 (predicted ≥ oracle in 82% of scenes, paired p=0.07). This is above a random-instance control (0.108, p=0.003) but not separable from the permuted control (0.188, p=0.58). In the 2048-token

Table 14: **Transplanted-object detection.** AUROC (transplant against every native labeled instance, pooled over official-val scenes; 61 floating / 62 floor scenes) and per-scene top-1 rate. err.: the predictor's error on the masked instance (CF protocol); sc.-mean and neighb.: static scores without a predictor; height: geometry only. One seed per arm except the rotary model (AUROC over three seeds, top-1 for seed 42). Our rows re-encode the edited scene; frozen-feature rows score the transplant against donor-scene tokens, which flags any encoding mismatch.

	AUROC				top-1 rate			
model	err.	sc.-mean	neighb.	height	err.	sc.-mean	neighb.	height
implausible: transplant floats 1–1.8 m above the floor								
reference (routed pair)	0.56	0.78	0.95	0.97	0.03	0.31	0.72	0.64
rotary model (reference + rotary predictor; AUROC n=3, top-1 seed 42)	**0.98**	0.71	0.94	0.97	**0.80**	0.13	0.66	0.64
pure SSL (pair recipe)	0.49	0.70	0.94	0.97	0.00	0.02	0.70	0.64
single token, α=1	0.66	0.85	0.97	0.97	0.08	0.30	0.80	0.64
SceneSplat frozen features, matched pred.	0.98	0.88	0.99	0.97	0.79	0.43	0.85	0.64
lifted DINOv2, no JEPA, matched pred.	1.00	0.94	1.00	0.97	0.97	0.56	1.00	0.64
SceneSplat self-supervised, matched pred.	0.98	0.70	0.97	0.97	0.80	0.26	0.72	0.64
plausible: transplant on the floor at its donor height								
reference (routed pair)	0.56	0.65	0.81	0.56	0.06	0.06	0.27	0.16
rotary model (reference + rotary predictor; AUROC n=3, top-1 seed 42)	**0.88**	0.60	0.80	0.56	**0.52**	0.10	0.19	0.16
pure SSL (pair recipe)	0.50	0.59	0.78	0.56	0.00	0.02	0.24	0.16
single token, α=1	0.58	0.58	0.84	0.56	0.06	0.08	0.29	0.16
SceneSplat frozen features, matched pred.	0.95	0.85	0.98	0.56	0.61	0.35	0.81	0.16
lifted DINOv2, no JEPA, matched pred.	0.99	0.95	1.00	0.56	0.94	0.68	0.98	0.16
lifted SigLIP2, no JEPA, matched pred.	0.99	0.96	0.99	0.56	0.94	0.71	0.98	0.16
SceneSplat self-supervised, matched pred.	0.96	0.68	0.91	0.56	0.60	0.16	0.47	0.16

Table 15: Retrieval class agreement: 62 official validation scenes $\times$ 3 seeded object-level masks, reference encoder, one 149-scene bank (target excluded). Mean $\pm$ s.d. over scenes (masks averaged within a scene); last column: paired difference, paired p, and the share of scenes in which the predicted branch is at least as high. The paired 90% interval for predicted $-$ oracle is $[-0.013, -0.001]$; the equivalence bound it implies (about 0.013) was chosen post hoc, so no equivalence is claimed. The 2048-token sibling gives 0.507 ± 0.116 vs. 0.511 ± 0.121 (p=0.23; interval $[-0.012, +0.002]$).

query embedding	class agreement	retrieval cosine	predicted $-$ this (paired)
predicted (predictor)	0.491 ± 0.118	0.810 ± 0.009	—
oracle (true, hidden)	0.498 ± 0.121	0.819 ± 0.011	-0.007, p=0.04, 42%
permuted within region	0.244 ± 0.107	0.819 ± 0.011	$+0.247$, $p=6\times10^{-28}$, 100%
unrelated groups	0.176 ± 0.107	0.819 ± 0.027	$+0.315$, $p=2\times10^{-27}$, 100%
scene mean	0.209 ± 0.175	0.269 ± 0.021	$+0.282$, $p=6\times10^{-16}$, 92%

sibling space the same vote separates from both controls (0.263 ± 0.326 vs. oracle 0.242 ± 0.326; permuted 0.183, p=0.006; random instance 0.113, $p=3\times10^{-4}$). The instance vote pools many small tokens at 4096 tokens and depends on the space; the retrieval claim rests on the token-level readout.

Retrieval across representations. With the matched predictors of the null battery and the same masks, SceneSplat's language-aligned features retrieve the right class more often than ours for the true token (0.70 vs. 0.50) and for the nearest-visible copy (0.55 vs. 0.43); relative to each space's oracle, the standardized SceneSplat predictor and ours reach 98% and 99%, and both beat the nearest-visible copy (+0.14 and +0.06, $p < 10^{-4}$ over scenes).

Unabridged text, retrieval agreement for every representation, retrieval at a removal void and retrieval-based completion renders: Extended results, Section 9 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

G ACTION-CONDITIONED LATENT WORLD MODEL

States, actions and re-encoding targets are defined in Appendix B; the reference space is primary (Table 16). Every space shares one protocol: ScanNet, 50 training and 15 validation scenes (592 and 178 synthesized transitions), identical PLYs and code, a depth-4 residual transformer trained with AdamW (3×10^{-4}, cosine schedule, 100 epochs, batch 32, or 16 in the 4096-token space) and SIGReg weight 0.05, the remove model starting at the identity. Composition and edit search use 30 held-out scenes: the 15 validation scenes and 15 labeled official-validation scenes.

Table 16: World-model results by latent space under the matched protocol. Remove: relative L2 and direction cosine on kept tokens. Add: additive marking, context tokens. Composition: two-step vs. one-step relative L2 on 30 held-out scenes (scenes beating identity). Search: exact edit-set recovery, 90 trials per K on 30 held-out scenes. ±: s.d. over three world-model seeds on one encoder checkpoint; the rotary cosine is the mean of three world-model seeds; other entries are single runs (sibling K=3: seed 42). —: not run.

	remove		add	composition	search	
space	rel. L2	cosine	rel. L2 / cosine	two / one step	K=2	K=3
reference	0.489 ± 0.017	0.861 ± 0.010	0.831 / 0.562	0.661 / 0.502 (30/30)	0.307 ± 0.006	0.141 ± 0.006
reference, eval. scenes excluded	0.499 ± 0.010	0.851 ± 0.013	—	—	0.333 ± 0.011	0.107 ± 0.017
rotary	0.593 ± 0.011	0.770	—	—	0.211 ± 0.011	0.104 ± 0.006
2048-token sibling	0.524 ± 0.008	0.829 ± 0.006	0.843 / 0.490	0.698 / 0.537 (30/30)	0.319 ± 0.017	0.14
base recipe	0.735	0.695	0.973 / 0.406	0.835 / 0.709 (29/30)	—	—

Remove, move and add. In the reference space three world-model seeds give remove relative L2 0.474, 0.507 and 0.485 (0.489 ± 0.017) and beat identity on 0.981 ± 0.004 of disturbed tokens; three encoder seeds give 0.481 ± 0.014. Beats-identity is 0.964 ± 0.007 in the sibling space, 0.910 in the base space and 0.92 in the rotary space. On InteriorGS, where a removal barely changes the other tokens, base-recipe world models trained the same way on an InteriorGS-pretrained encoder yield the identity solution. Moves from 0.5 to 6× the object's extent lower the reference's all-token relative L2 from 0.691 to 0.591, and both routed-pair spaces have fitted midpoints below the smallest magnitude tested (reference: x_0 at 0 in every scene-level bootstrap resample). The base space's ScanNet midpoint is 0.94× (95% CI $[0.81, 1.10]$), against 2.40× for a base-recipe encoder pretrained on InteriorGS (n=1; fit-sensitive); with base-recipe encoders pretrained on each corpus, ScanNet edits perturb the latent about 8× more than InteriorGS edits, consistent with a threshold that depends on the corpus. Add is same-scene re-insertion: the input is the scene without the object together with the object's tokens encoded on their own and marked as arriving, and the target is the original full-scene encoding. With additive marking it beats identity on 72% of context tokens in the reference space; insertion is harder than removal because it predicts the scene's change from the object's tokens encoded without their context.

Only the learned model transfers edit structure. On the 30 held-out scenes, non-learned transition baselines do no better than identity: nearest-neighbor transition relative L2 1.015 and linear mean-delta 1.000 in the reference space (sibling 1.019 and 1.000; base space on context tokens: world model 0.709, nearest neighbor 1.036, linear 1.0).

Composition and edit search. The remove model is applied twice (the second removal from the *predicted* state) and compared with the re-encoded doubly edited scene on 30 held-out scenes (15 world-model validation and 15 labeled official-validation scenes, none in the training split; results in

Table 16). Edit search follows a protocol pre-registered before the first run: on the same 30 scenes, the six largest instances per scene by token count (walls and floor included, no semantic filter) each admit five actions (remove; move $\pm x$ or $\pm y$ by $3\times$ its extent); K random edits on distinct objects define a goal latent through the exact simulator (drop removed groups, shift moved centers, re-encode); three goals per scene give 90 trials per K. The search rolls the remove and move models over every candidate set (375 at $K{=}2$, 2500 at $K{=}3$) and picks the set whose predicted end state is closest to the goal over *context* tokens only. Baselines are a random set, identity dynamics, a greedy locality heuristic without the world model, the same heuristic reading remove versus move off the goal's zeroed slots, and the exact simulator as an oracle. For the first world-model seed, the search recovers the exact set in 0.311 of trials at $K{=}2$ (95% scene-bootstrap CI $[0.22, 0.40]$; chance 0.0027) and 0.144 at $K{=}3$ ($[0.07, 0.22]$; chance 0.0004), and executing the recovered plan leaves 0.20 and 0.34 of the unedited scene's distance to the goal, against 1.42 and 1.08 for random plans (Figure 8). One-sided exact McNemar tests paired by trial (seed 42) give wins/losses at $K{=}2$ of $28/0$ against random and greedy ($p{=}4\times10^{-9}$), $27/5$ against greedy with zeroed slots ($p{=}6\times10^{-5}$) and $25/2$ against identity dynamics ($p{=}3\times10^{-6}$); scene-level sign-permutation tests, which respect the three goals per scene, give $p < 0.003$ against random, greedy and greedy with zeroed slots for every seed at both horizons. The oracle recovers 1.00 at both horizons, so the remaining gap is world-model error. Object-set recovery does not measure action recovery: identity dynamics reads removed objects off their zeroed slots and scores higher on it.

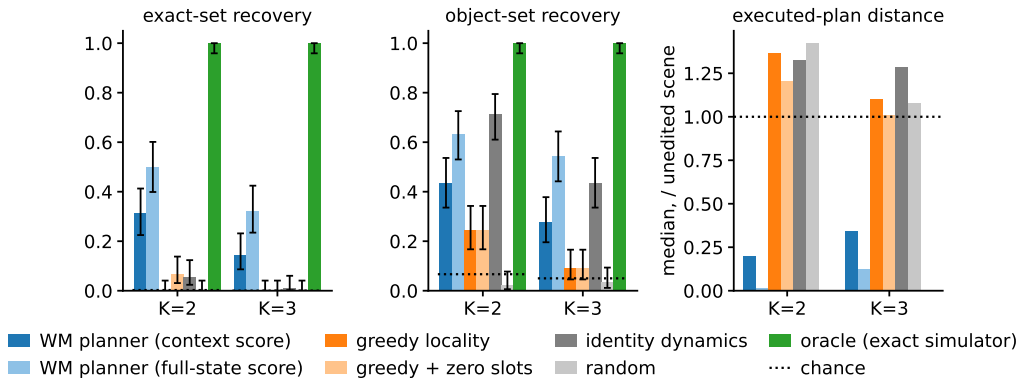
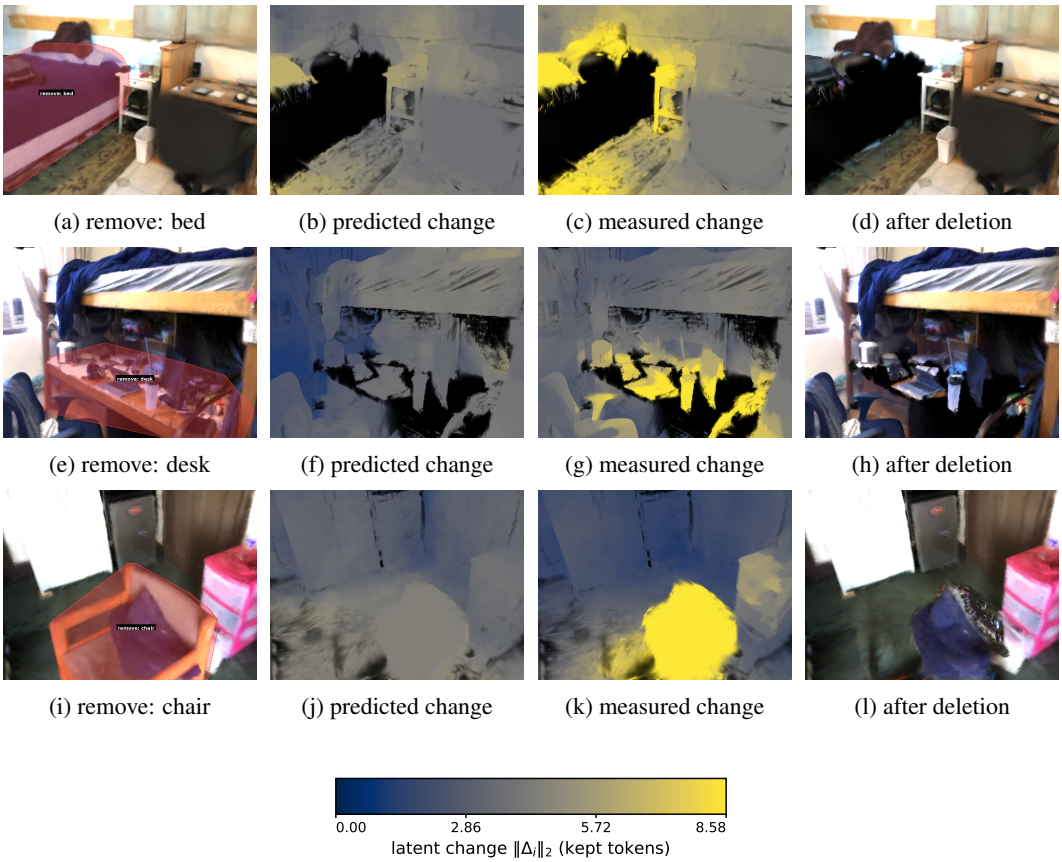


Figure 8: **Edit search in latent space** (reference space, 90 trials per K). Exact-set and object-set recovery, and the executed plan's median distance to the goal relative to the unedited scene (dotted: chance for recovery, 1.0 for distance). Under the pre-registered context-only score (dark blue) the search leads every non-oracle baseline on exact-set recovery and executed-plan distance at both horizons; identity dynamics scores higher on object-set recovery by reading the zeroed slots. The exact simulator (green) shows the task is identifiable. Light blue: a secondary full-state score that also reads the object slots, reported for completeness; the pre-registered score is context-only. Error bars: 95% Wilson intervals.

Fixed-frame decomposition of edit responses. For the world model's 15 validation scenes and the reference encoder, each transition's change Δ_A (standard re-encoding) is compared with Δ_B (the edited scene encoded in the pre-edit coordinate frame: the fixed-frame change) and Δ_C (the unedited scene encoded in the post-edit frame: the normalization-only change); values are per-scene means, then the mean over scenes ($n{=}15$). This evaluates encoder targets, not world-model predictions: in the rest of the scene the standard change follows the normalization-only change, whereas the moved object's change follows the fixed-frame change (Table 17). The components are not additive, and cosines are not fractions of variance explained.

World-model predictions against the fixed-frame change. Applied to the predicted change $\hat{\Delta}$ (one reference-space checkpoint per action and move scale; the same 15 scenes; $\pm$ over scenes), the comparison shows that on rest-of-scene tokens after removal the prediction aligns with the standard target (0.654) about as much as with the normalization-only change (0.649). After projecting both $\hat{\Delta}$ and Δ_B off Δ_C for each token, their alignment is 0.111 ± 0.048, positive in 15 of 15 scenes.



(a) remove: bed (b) predicted change (c) measured change (d) after deletion

(e) remove: desk (f) predicted change (g) measured change (h) after deletion

(i) remove: chair (j) predicted change (k) measured change (l) after deletion

Figure 9: **Predicted and measured latent change after an object removal** (reference encoder and its remove model; one real capture pose per row): the object outlined; kept Gaussians colored by the predicted change of their token, $\|\hat{\Delta}\|$; the same for the change measured by re-encoding (one color scale shared by all rows, bottom); the scene after deletion. Relative L2 / direction cosine / beats-identity / Spearman between maps: bed (74 tokens) $0.56/0.80/100\%/0.76$; desk (38 tokens) $1.54/-0.23/7\%/0.12$; chair (31 tokens) $1.93/-0.21/5\%/0.49$. On the desk and chair rows the prediction is diffuse where the measured change concentrates at the void. Maps include coordinate-normalization effects.

Table 17: Where edit responses come from (reference encoder, 15 validation scenes). Rest of scene: kept tokens other than the edited object; moved object: its own tokens.

	rest of scene					moved object	
edit	$\|\Delta_A\|$	$\|\Delta_B\|$	$\|\Delta_C\|$	$\cos(\Delta_A, \Delta_B)$	$\cos(\Delta_A, \Delta_C)$	$\cos(\Delta_A, \Delta_B)$	$\cos(\Delta_A, \Delta_C)$
remove	9.08	2.68	8.27	0.44	0.92	—	—
move 0.5×	5.94	2.87	4.86	0.52	0.80	0.88	0.09
move 1×	9.30	4.61	7.97	0.49	0.80	0.82	0.21
move 2×	13.45	6.08	12.00	0.43	0.81	0.75	0.29
move 4×	18.05	6.70	16.86	0.34	0.84	0.71	0.34

The pre-registered criterion required a mean of at least 0.30 and positive values in at least 13 of 15 scenes: the sign condition is met, the mean condition is not, so the criterion is missed. Moves show the same pattern (projected alignment rising from 0.090 ± 0.038 at $0.5\times$ to 0.177 ± 0.034 at $4\times$, positive in every scene at every scale): the prediction carries a small but consistently positive component of the fixed-frame change beyond the normalization direction. This diagnoses models trained on standard re-encoding targets. **Trained on fixed-frame targets** for 200 epochs in a follow-up, the world model reaches a projected alignment of 0.375 ± 0.014 (s.d. over five seeds), above the

same 0.30 threshold in every seed and positive in all 15 scenes; its direct alignment with the fixed-frame target is 0.418 ± 0.017 and its relative L2 0.811 ± 0.010. At the originally pre-registered 100 epochs, two of three seeds meet the registered criterion; the third had not yet left an early plateau.

More training scenes. Enlarging the synthesized training set from 50 to 385 scenes, with the same encoder, protocol and 15 validation scenes, lowers relative L2 steadily (Table 18); the larger sets also win at fewer gradient steps.

Table 18: World-model data scale (reference space, remove; three world-model seeds each; 100 epochs). Larger sets also take more gradient steps, yet win at fewer: 150 scenes after 30 epochs (53k samples) reach 0.422 and 385 scenes after 10 epochs (46k) reach 0.387, against 0.489 for 50 scenes after 100 epochs (59k); early values are means over the three seeds from the validation logs. No training scan shares a room with a validation scan.

training scenes	transitions	relative L2	direction cosine	beats identity
50	592	0.489 ± 0.017	0.861 ± 0.010	98%
150	1777	0.369 ± 0.003	0.927 ± 0.001	99.9%
385	4551	0.277 ± 0.001	0.961 ± 0.0002	99.99%

Unabridged text, edit-search floor plans, action decoding from the latent change, the replace-marking variant of add and the locality of re-encoding changes: Extended results, Section 10 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

H TRANSFER BOUNDARY TO REAL RESCANS

The world model is trained on synthesized edits of single fits; this section measures what separates those transitions from real re-captures. The 3RScan, Remove360 and PanopticSports measurements were made in the base-recipe space; on 3RScan, real-frame-distilled encoders show floors of 0.25–0.47, and Table 19 measures the floor in the reference, image-free and single-token spaces, so the fit-noise floor appears in every space we tested.

3RScan rescans. On 330 real (original scan, rescan) pairs of 3RScan scenes with pre-fit MCMC Gaussians and ground-truth object-level changes, two independently fitted reconstructions of the same unchanged region differ by a matched-token latent drift of 0.17–0.50 (median), as large as most edit-induced change. The base-recipe remove world model (512 tokens, evaluated with the encoder seed it was trained with), applied zero-shot to 25 of these pairs with exact ground-truth edit masks, beats identity on 11 of the 25; the effect depends on the input, but 99 contiguous mask-placement permutations per pair show that it is not localized to the true edit (Fisher $p \approx 0.47$ for wins).

The floor on ScanNet re-captures. ScanNet's repeat scans of one space are independent 3DGS fits by the pipeline the encoder was trained on and are held out from the controlled splits, which use each room's first scan only. We register each of 452 (first scan, rescan) pairs by a yaw-and-translation fit on the 2048 group centers (11 pairs with residual above 0.30 m excluded), match groups by mutual nearest center within 0.25 m and report the frozen encoder's per-group drift $1 - \cos$ (Table 19). Drift is the same where majority labels agree and disagree, consistent with fit noise rather than semantic change, and is a large fraction of the random-token distance.

What lowers the floor. On Remove360 [28] (one object physically removed; our own COLMAP and MCMC pipeline), independent fits give a floor of 0.14–0.26 and warm-initialized refits, which run 3k unconstrained MCMC steps from the before-fit's converged parameters, give 0.138–0.235. Frozen local refits [35; 48], which optimize only the Gaussians near the change, lower the floor to 2.4×10^{-4}–3.0×10^{-3} and leave no residual Gaussians of the removed object inside the frozen region: freezing the unchanged region removes the floor, and a shared initialization does not. On temporally consistent fits of PanopticSports sequences [54; 34], the background tokens' geometry-channel floor is 7×10^{-7}–6×10^{-5} across frame gaps of 1–30.

Table 19: **Same-room re-capture drift** on 441 ScanNet space pairs (per-group $1 - \cos$ of the frozen encoder between two independent fits of one room after registration; medians over pairs). Random token: the same distance to a random group of the other scan.

encoder	drift	random token	labels agree	labels disagree
reference (routed pair, 4096 tokens)	0.71	0.97	0.71	0.71
pure SSL (same recipe, no image teachers)	0.89	1.00	0.89	0.89
single token, 2048 groups (visregSingle2048)	0.67	0.99	0.68	0.70

The world model once the floor is removed. On (before, frozen-after) pairs for all three Remove360 scenes, the base-recipe world model favors identity in every scene ($0/3$; identity drift 0.08–0.13 vs. world-model drift 0.10–0.23). Its predicted disturbance norms separate Remove360 inputs from its ScanNet-validation distribution (Kolmogorov–Smirnov $D{=}0.364$; tokens pooled over three scenes), consistent with a shift from synthetic to real transitions, although the test does not identify the cause. Our working explanation is magnitude calibration: a frozen-refit removal changes the rest of a real scene little, while a model trained only on synthesized deletions predicts larger disturbances. Base-recipe world models trained this way beat identity on synthesized transitions (relative L2 0.735 under our protocol; reference space 0.489, Table 16), so the gap points to the mismatch between synthetic and real transitions, not to an inability to predict edits in-distribution.

Refit fidelity on real rescans. Frozen refits of real ScanNet rescan pairs were evaluated under a gate fixed in advance: median PSNR ≥ 20 dB on held-out views of the after-capture. Under rigid registration the pre-registered primary rule passes 1 of 86 held-out pairs (raw median 9.43 dB) and the secondary rule 0 of 86; 42% of these pairs do not reach 20 dB even with their own same-capture reference. On ten development pairs, a learned registration gauge raises the median from 9.49 to 13.48 dB and passes one pair; a bar fixed before these runs required at least four passing pairs, so the 86-pair sweep was not run.

Unabridged text, change detection across token spaces, render space and raw parameters, and the refit-fidelity diagnostics: Extended results, Section 11 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).

I Extended related work

Gaussian-splat representations and probe protocol. SceneSplat [4] distills vision–language features into scene-level splats and ShapeSplat's Gaussian-MAE [6] reconstructs masked Gaussian parameters; we instead predict contextual embeddings of masked scene regions. Farganus et al. [57] compare point- and graph-based models for Gaussian-scene classification and evaluate their representations through linear probing and clustering; our study measures prediction of hidden scene representations and specified edit effects. SceneSplat-Bench [5] classifies splat features zero-shot by text embedding with kNN label transfer and foreground exclusion, so we do not quote its published SceneSplat number (≈ 0.35 zero-shot f-mIoU) and re-measure every SceneSplat value under our linear-probe harness.

Concurrent work. Gaussian-JEPA [43] predicts held-out object-level Gaussian blocks with an EMA teacher on ShapeSplat objects; ours is scene-level, replaces the momentum teacher with a distributional regularizer, and is evaluated by probing, prediction margins over matched nulls, retrieval and a latent edit model. SR-JEPA [40] queries a frozen predictive pathway at a removed object's centroid on point clouds, with EMA targets and no image teachers; its donor-context and randomized-path floors play the role of our nulls, without shuffled- or mean-context terms or an edit model. 4DGS-WAM [45] is an action-conditioned, reconstruction-driven world model on Gaussian splats for outdoor driving; our world model predicts a frozen latent learned without class labels and renders nothing. 4DGS-JEPA [42] studies temporally compositional prediction of dynamic Gaussian states with selective geometry decoding; we learn masked representations from captured static scenes and train a separate latent model for specified object edits and edit search. Lentsch et al. [49] find that the instance signal of a frozen self-supervised LiDAR point transformer is carried predom-

inantly by its rotary position encoding; our rotary result is the predictor-side counterpart on indoor splats.

Trade-offs and world models. Our recognition–prediction trade-off is an instance of a trade-off documented by IWM [19] and Guillotine Regularization [37], here with a Concerto-style [36] image teacher as the controlled variable. Action-conditioned latent predictors include DINO-WM [20] and LeWM [21], whose prediction-plus-SIGReg loss we adopt; Gaussian-splat world models [22; 23; 24] learn from robot or video trajectories, whereas ours amortizes a known map, the encoder's re-encoding of an edited scene.

The full related-work discussion: Extended results, Section 12 (`https://anonymous.4open.science/r/iclr-anon-3dgs-wm-F303`).