BRIDGE: TURNING FROZEN FOUNDATION-MODEL FEATURES INTO EXPLICIT STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained generative models hold information that is useful for tasks they were never trained to solve. Getting at it is the hard part: a downstream model has to select the relevant intermediate features and still respect the dependencies among its own outputs. We present BRIDGE, a lightweight architecture that reads the intermediate activations of a frozen model and turns them into the different representation a task requires. A small set of capture sites is compressed into a queryable state of fixed size. Each output element forms its own query from its current state and geometry, a shared global state lets elements exchange information before they are updated, and the update itself is whatever the task needs. The reader is indifferent to which model supplied the activations and to the form of the output head.

We instantiate BRIDGE for sparse-view 3D surface completion from a frozen video model: it predicts a fixed population of oriented surface points and carries their positions and normals jointly with conditional flow matching, so that each local prediction sees its own geometry and scene-level information together. The same reader also runs in the opposite direction. Trained on a frozen vision–language model's activations and then frozen, it becomes a differentiable geometric critic whose error is backpropagated into a low-rank adapter of the language model. With the critic removed, the adapted model raises its strict successes on held-out objects it had failed from 10 to between 17 and 20 of 60 under an unchanged criterion, while an untrained critic of the same architecture yields no gain. A reader of intermediate representations can thus both extract structure from a frozen model and feed structure back into it.

1 INTRODUCTION

A video generator conditioned on a handful of photographs has to settle many questions that its training objective never names. It decides how a wall continues behind a chair, how far apart two objects stand, and which silhouette belongs to which surface. Those decisions are already present in the model's intermediate activations, before any pixel is written. For a downstream 3D task this is an attractive source of evidence: the prior has been paid for once, on a corpus far larger than any reconstruction dataset, and it can be read without completing a generation and inverting it afterwards.

Turning intermediate activations into a usable interface is the actual problem. The state of a large video model is both wide and long. One transformer layer at one denoising step holds tens of thousands of feature vectors of width 5376, and useful evidence is spread over several layers and several steps. A downstream network that ingests all of it inherits the cost of the generator and loses the advantage of reading it early. The activations are also indexed the wrong way for geometry: they live on a lattice of video frames and image patches, while a surface has to be produced in metric coordinates. A downstream model therefore has to choose what to read, and the right choice depends on where in space it is currently predicting.

BRIDGE makes that choice part of the architecture, in three stages. A *Compressor* reads activations at a fixed, small set of *capture sites*, each a chosen pair of transformer layer and denoising step, compresses local neighborhoods of the feature lattice, and resamples them into a short sequence of learned summaries. What remains is a *Memory* of fixed size that any consumer can query at low cost. A *Reader* does the querying for the task output: each output element forms a probe from its own

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

current state, retrieves evidence from Memory, and reconciles the returns into one condition vector. A *Head* consumes that condition and advances the output state. Only the Reader touches both sides, through a fixed-width probe and one attention read per site, so the source model and the output head are interchangeable parts rather than architectural commitments.

We instantiate this interface for sparse-view 3D surface completion. From one to nine source images, BRIDGE produces a fixed population of oriented points: a position and a unit normal per point. Points are carried by a conditional flow that we integrate with an explicit Euler scheme. At each evaluation, nearby points are collected into local groups, each group forms one query from its own shape, location, size, and time, and the retrieved evidence conditions the velocity of every point in that group. The compressed activations are shared by all groups and supply scene-level context to every local prediction, while learned latents inside a group circulate information among its points. Local structure and global agreement come out of the same update, and no point is decoded in isolation.

The three stages hold 11,692,042 trainable parameters between them, split almost evenly, and the source model stays frozen. The instance we measure reaches 60.98% mean F-score at a 3.46 cm threshold on validation scenes it was not trained on and 83.54% on training scenes, where the strongest case reaches 93.31% precision, 99.81% coverage, and 1.17 cm Chamfer-L_1. Section 4.1 states two properties of the container behind those numbers: it is a permuted variant of the published attention weights, and it was read by the preceding instance of the same three stages. Controlled interventions locate the performance: replacing Memory with that of a different scene costs 8.38 points of precision, removing the learned spectral mixture 5.55, and removing the consensus between evidence and geometry 5.81 points of precision and 6.65 of normal accuracy under a matched adaptation budget.

The Reader also works in reverse. On a frozen vision–language model that is asked to rebuild a composed solid as a program, we train the Reader and Head on the model's own activations, freeze them, and backpropagate their geometric error into a low-rank adapter of the language model (Section 4.4). After the critic is removed, the adapted model solves 17 to 20 of 60 held-out objects it had failed, against 10 before adaptation, and an untrained critic of the same architecture brings no gain on average.

Our contributions are:

- **A readout that turns frozen intermediate representations into a queryable Memory.** A small number of capture sites, a complete local spectral basis with a learned contraction, and learned resampling reduce a wide activation field to a fixed-size state that the task output queries with its own evolving geometry. Neither the identity of the source model nor the form of the output head enters the construction.
- **A jointly updated oriented-point flow.** Output elements are grouped by current position, share a scene-level state, and exchange information through peer latents. A local prediction therefore carries its own geometry and the scene context together.
- **An 11.7M-parameter instantiation with measured feature dependence.** Interventions on the frozen model and a matched-budget factorial study isolate the contribution of each part of the interface, and a function-preserving depth construction gives the enlarged family at least the same representational capacity.
- **A frozen Reader as a differentiable critic for a language model.** The same Reader, trained on a vision–language model's activations and frozen, turns geometric error into a training signal for that model and repairs its held-out failures without being present at test time.

2 Related work

Video priors for explicit 3D. Several recent systems recruit a video generator for reconstruction from sparse views. ReconX conditions generation on an image-derived point cloud and then optimizes a Gaussian scene from the generated frames (Liu et al., 2026). Lyra distills a video diffusion model into a Gaussian decoder that consumes denoised video latents with camera-ray embeddings (Bahmani et al., 2026). SpatialCrafter combines video latents, monocular depth, and camera information in a feed-forward Gaussian reconstructor (Zhang et al., 2025). All define their interface at the generator's output or its final latent. BRIDGE defines it over intermediate activations taken at selected layers

and denoising steps, and lets the downstream task decide which to read. MapAnything (Keetha et al., 2026) provides the image-only geometric initializer and the coordinate frame for our instantiation.

Surface flow and communication between outputs. Surflo compresses unposed views into a fixed-size global state and transports independently decoded oriented points onto a surface with flow matching, coupling nearby points at inference through a photometric guidance loss (Guédon et al., 2026). PC^2 projects image features onto the current noisy point cloud at every denoising step (Melas-Kyriazi et al., 2023). PointInfinity circulates information between a variable number of point tokens and a fixed-size latent stream (Huang et al., 2024). BRIDGE keeps the evolving geometry as the probe and makes communication part of the learned update: groups are rebuilt from current positions at every velocity evaluation, each group probes the shared Memory, and peer latents exchange information inside a group during the same forward pass.

Reading intermediate representations. Diffusion Hyperfeatures aggregates activations across layers and denoising steps into a descriptor field for semantic correspondence (Luo et al., 2023). Readout Guidance trains small readout heads on a frozen image diffusion model's features and uses their predictions to guide sampling (Luo et al., 2024); Diff3F distills diffusion features onto untextured shapes (Dutt et al., 2024), and 3DiffTection makes them geometry-aware for 3D detection (Xu et al., 2024). These readouts decode a fixed map per image. The BRIDGE Reader is instead queried by the output's own evolving state at every integration step, it drives an explicit 3D output, and once frozen it trains the source model's weights rather than steering its sampling. PointNet-style pooling (Qi et al., 2017) over the current point group produces that query, and compression precedes any geometry-dependent access, so the wide projection is paid once per scene. We train the Head with Flow Matching (Lipman et al., 2023) and grow capacity by a function-preserving construction in the spirit of Net2DeeperNet (Chen et al., 2016); Appendix A adds the remaining comparisons.

3 ARCHITECTURE

A frozen generative model holds its evidence on one index system, a lattice of video frames and image patches. A task produces its output on another, here metric space. BRIDGE is the interface between them, built from three stages of almost equal size (Table 3 in Appendix B). The **Compressor** sees only the source: it reduces the activation field of each capture site to a fixed-size *Memory*, once per request. The **Reader** is the crossing point: each output element forms a *probe* from its own current state, addresses Memory with it, and reconciles what comes back into one *condition* vector. The **Head** sees only the task. Only the Reader touches both sides, through a probe of width 256 and one attention read per site; neither the source width nor the output population size appears in that interface, so the Compressor can change with the source model and the Head with the task without either seeing the other.

We instantiate the three stages for sparse-view 3D surface completion. The output is a population of *particles* with state $Z_t = [X_t, D_t] \in \mathbb{R}^{N\times 6}$, where X_t holds positions in meters, D_t holds unconstrained direction vectors that become unit normals at the end, and $t \in [0, 1]$ is integration time. An initializer supplies *birth anchors*, the centers around which the initial particles are drawn, with their per-anchor areas and the coordinate frame. The population is allocated once and fixed for the whole rollout: no particle is later added, removed, resampled, or reordered.

3.1 COMPRESSOR: THE ACTIVATION FIELD BECOMES MEMORY

A *capture site* is a chosen pair of transformer layer and denoising step in the frozen generator, MiniMax-H3 (MiniMax-AI, 2026). We use twelve sites, listed in Appendix B. A capture holds a packed sequence of feature vectors over all of the generator's input streams; we keep the vectors of the video being generated and restore their native order,

$$F_s \in \mathbb{R}^{T_s\times H_s\times W_s\times 5376}, \tag{1}$$

where s indexes the site, T_s indexes video time, and H_s, W_s index spatial patches of the generator's latent grid.

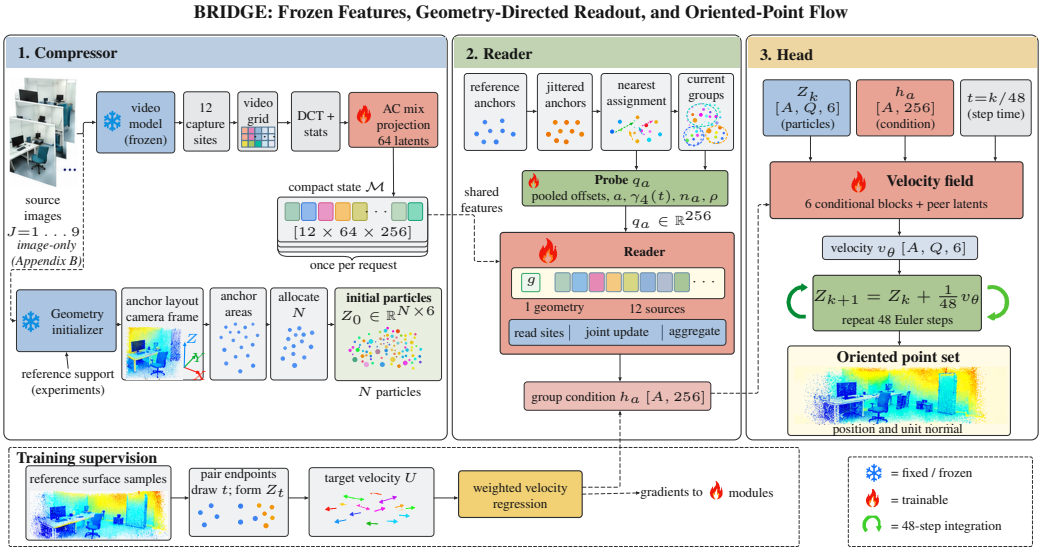


Figure 1: Overview. The Compressor encodes the frozen activation field once per request into Memory. At every velocity evaluation the Reader forms one probe per particle group, addresses Memory, and emits a condition; the Head turns that condition into a velocity for every particle of the group. Geometry initialization uses reference support in the reported experiments; the dashed image-only route is described in Appendix B. Training pairs the endpoints and regresses $U = Y - Z_0$. Office and point-cloud insets are schematic.

Window spectrum. For each site we accumulate the channelwise mean $\mu_s \in \mathbb{R}^{5376}$ over the complete field and one centered root-mean-square scalar σ_s. At each time index we partition the $H_s \times W_s$ grid into non-overlapping *windows* of 4×4 cells. For a window w with $m_w \le 16$ valid cells, let $F_{s,w} \in \mathbb{R}^{m_w \times 5376}$ collect its vectors and $B_w \in \mathbb{R}^{16\times m_w}$ hold the orthonormal modes of the two-dimensional discrete cosine transform over the window's own valid cells, acting across cells within each channel. The normalized coefficients are

$$C_{s,w} = B_w \left(F_{s,w} - \mathbf{1}\mu_s^\top\right)/\sigma_s \in \mathbb{R}^{16\times5376}. \tag{2}$$

An interior window has 16 cells and 16 modes, so Equation 2 is a complete change of basis and discards nothing: we measure $\max|B^\top B - I| = 9.5 \times 10^{-7}$ and a relative FP32 round-trip error of 4.1×10^{-7}. Row 0 is the spatially constant component, rows 1 to 15 the spatial variation.

Variation mixer. Contracting the fifteen variation rows is the only place information is lost. A small network predicts four sets of signed weights $\alpha_{s,w,j,\cdot} \in \mathbb{R}^{15}$ from a learned 32-dimensional site identifier, a four-band Fourier encoding of the normalized window center in (time, row, column) of width 27, the sixteen per-mode log amplitudes, and the support fraction $m_w/16$; these 76 values pass through a $76 \to 64 \to 60$ multilayer perceptron, $\tanh$, and ℓ_2 normalization within each set. The window becomes five rows,

$$P_{s,w} = \left[C_{s,w,0};\ \left\{ \textstyle\sum_{r=1}^{15} \alpha_{s,w,j,r}\, C_{s,w,r} \right\}_{j=1}^{4} \right] \in \mathbb{R}^{5\times5376}. \tag{3}$$

Four learned directions rather than one matter here: a single mixed row leaves one reachable variation direction out of the fifteen the basis provides. A shared $5376 \to 256$ projection with a SiLU activation maps the rows to width 256, and we add embeddings for window position, site, support, and row role. Site s yields $K_s = 5\,T_s\lceil H_s/4\rceil\lceil W_s/4\rceil$ local tokens, every native temporal slice surviving; on the measured lattice of a square request this is 1,998 windows and $K_s = 9{,}990$. Seeing five rows per window instead of sixteen cells cuts the wide projection's matrix-product term by 3.2 (Appendix D).

Resampler. K_s varies with the aspect ratio and length of the generated video, so a consumer that queried it directly would face a variable-length sequence at every step. Sixty-four learned latents

per site read all K_s local tokens with four heads of width 64, followed by self-attention and a gated SwiGLU update, producing $M_s \in \mathbb{R}^{64\times256}$. The latents form four temporal bands of sixteen content slots, and three additive biases shape the read: a *band prior*, one learned center and width per band along video time; a *place prior*, one learned center and width per latent over the window's normalized spatial position; and a log-support term. Every latent keeps access to every token, so the priors bias without restricting. The place prior binds a latent to a region of the image plane, and a learned center is scene-independent where an absolute coordinate is only memorized. A separate branch projects the site's mean and scale statistics, removed by centering and normalization, into one summary vector. Memory is

$$\mathcal{M} = \{(M_s, b_s)\}_{s=1}^{12}, \qquad M_s \in \mathbb{R}^{64\times256}, \quad b_s \in \mathbb{R}^{256}, \tag{4}$$

fixed in size regardless of video resolution or length, and built once per request.

3.2 Reader: each element addresses Memory

Grouping. Reading is done per group. Grouping anchors come from the same immutable reference base as the birth anchors and never from the current output: at every velocity evaluation we redraw a per-anchor offset uniformly in a volume ball of radius at most $0.05\,\mathrm{m}$ and assign every particle to its nearest jittered anchor. Every particle belongs to exactly one anchor, including far outliers; the reach $r = 0.30\,\mathrm{m}$ sets the normalization and the area bookkeeping, not an admission cutoff. Jitter does not accumulate and the assignment is rebuilt each time, so groups follow the particles as the surface forms.

Probe. For particle i in the group at anchor a, a shared PointNet-style multilayer perceptron (Qi et al., 2017) maps the local offset $(x_i - a)/r$ through widths $3 \to 64 \to 128$ with SiLU activations to $e_i \in \mathbb{R}^{128}$. Masked mean and maximum pooling over the group's real particles, a closed-form shape descriptor $g_a \in \mathbb{R}^9$, the anchor place code, the integration time, the group size, and the requested density give

$$q_a = \mathrm{LN}\left(f_\theta\left[\mathrm{mean}_{i\in\mathcal{G}_a}\, e_i,\ \max_{i\in\mathcal{G}_a} e_i,\ g_a,\ \gamma_8^{32}(a),\ \gamma_4(t),\ \frac{\log(1+n_a)}{10},\ \log\frac{\rho}{1000}\right]\right) \in \mathbb{R}^{256}, \tag{5}$$

where $n_a = |\mathcal{G}_a|$, f_θ is a $327 \to 256 \to 256$ multilayer perceptron, LN is layer normalization, γ_b^p concatenates its input with sine and cosine features at frequencies $2^k\pi/p$ for $k = 0, \ldots, b-1$, and ρ is the requested number of points per square meter.

Two inputs deserve comment. The place code $\gamma_8^{32}(a)$ resolves $0.50\,\mathrm{m}$ at its finest band, deliberately coarser than the Head's $1.56\,\mathrm{cm}$: a probe says which part of the scene is asking, and since scenes occupy disjoint coordinate ranges, a centimeter-resolution absolute code can only memorize what sits at a given coordinate. The shape descriptor holds the normalized eigenvalues of the group's position covariance, the largest eigenvalue $\lambda_{\max}$ of the mean outer product of unit directions, the mean unit direction with its norm $\|\bar{n}\|$, and the extent in units of r. Because $(-n)(-n)^\top = nn^\top$, $\lambda_{\max}$ is blind to normal sign while $\|\bar{n}\|$ is not, so the pair separates a single sheet, $(1, 1)$, from two opposed sheets sharing one anchor, $(1, 0)$, from a scattered group, both below 1. The second case is the failure mode of purely Euclidean grouping and it is common (Appendix C).

Slot read and consensus. The probe reads every site, and the twelve results are reconciled with it before any reaches the Head. A learned summary query first reads each M_s and adds the site's summary vector to form b_s. Conditioned on q_a, on time, and on these summaries, the probe emits sixteen read slots per site; the concatenated read is split into four groups of four slots, each projected to width 256, and the four projections summed into that site's state. Splitting the projection raises its rank limit from 256 to 1024 at an unchanged parameter count, and the state stays 256 wide so nothing downstream widens.

The twelve site states and the probe state form a joint $[13, 256]$ state that two four-head self-attention and SwiGLU layers update together. This is the *consensus*: evidence from different sites and the geometry that requested it agree here, before anything is committed. The updated probe state then attends over the twelve updated site states, and a residual $256 \to 512 \to 256$ SiLU projection produces the condition

$$h_a = \mathcal{A}_\theta(q_a, \mathcal{M}, t) \in \mathbb{R}^{256}. \tag{6}$$

Every group derives its own h_a from the same Memory. One fixed-size attention read per site is what scene-level context costs a local prediction. Appendix B gives the intermediate operations and shapes.

3.3 HEAD: THE OUTPUT STATE ADVANCES

Given h_a, the Head updates the group's particle features through six conditional residual feed-forward blocks of width 256. World-frame positions and raw directions are Fourier-encoded on separate bands, positions with eight bands at unit period and directions with three bands at period 2, and a linear map takes the concatenated width 72 to width 256; separate bands prevent the direction channels from aliasing (Appendix C). Only the three raw position channels are divided by 16; Fourier phases stay metric. The condition and a 64-dimensional time embedding are concatenated to width 320, projected to 256, and set the scale and shift of adaptive layer normalization in every block, each of which applies a single-path SiLU feed-forward update with hidden width 512.

Blocks contain no attention. All communication between particles happens at two *peer exchanges*, before blocks 2 and 4: 128 learned peer latents read the group's current particle features and write the aggregated result back to every valid particle through a second attention pass. Points in a group therefore see each other twice while they are updated, and the shared condition keeps distant groups consistent. A final linear map produces the velocity in $\mathbb{R}^6$. Padding is sanitized before the Fourier encoding and zeroed at the output.

Training. Conditional flow matching (Lipman et al., 2023) regresses the velocity on straight paths between a sampled initial state Z_0 and a reference endpoint Y,

$$Z_t = (1-t)Z_0 + tY, \qquad U = Y - Z_0, \qquad t \sim \mathcal{U}[0,1), \tag{7}$$

where Y holds reference positions and unit normals sampled from the supervision surface. Initial positions are drawn with $0.1\,\mathrm{m}$ Gaussian noise around jittered birth anchors and initial direction components with standard deviation $1/\sqrt{3}$; XYZ endpoints are paired by minibatch Sinkhorn matching followed by a greedy bijection, blended with independent noise, and never paired across birth domains (Appendix C). Probes and groups are always built from the current interpolated state Z_t. Writing $v_i = (v_i^x, v_i^d)$ for the predicted velocity and $U_i = (U_i^x, U_i^d)$ for its target,

$$\mathcal{L} = \frac{25}{3Nr^2}\sum_{i=1}^{N}\|v_i^x - U_i^x\|_2^2 + \frac{1}{3N}\sum_{i=1}^{N}\left\|v_i^d - U_i^d\right\|_2^2. \tag{8}$$

Complete groups are batched into padded packs covering the whole population, and every real particle enters with the same scene-level denominator N, so the objective is invariant to how the packs are cut (Appendix D). Each optimizer step accumulates eight equally weighted scene computations. Compressor, Reader, and Head train jointly, and the source model stays frozen.

Inference. The default solver is a 48-step explicit Euler integration,

$$Z_{k+1} = Z_k + \frac{1}{48}\, v_\theta\big(Z_k, \tfrac{k}{48}; \{h_a(Z_k, \tfrac{k}{48})\}\big), \qquad k = 0, \ldots, 47, \tag{9}$$

which rebuilds the grouping, probes, site reads, and conditions from the updated state at every step, while Memory is built once and reused across all 48. At the endpoint each raw direction d_i is decoded as the unit normal $n_i = d_i/\|d_i\|$.

Capacity. The trained instance holds 11,692,042 parameters, which keeps all three stages in one graph on a single device at full population. Depth can be added without leaving the trained function behind: fixing the Compressor, state width, grouping rule, peer-exchange locations, and output map, the family $\mathcal{F}_d$ of d conditional blocks satisfies $\mathcal{F}_d \subseteq \mathcal{F}_{d+k}$, so

$$\inf_{f \in \mathcal{F}_{d+k}} \mathcal{R}(f) \;\le\; \inf_{f \in \mathcal{F}_d} \mathcal{R}(f) \tag{10}$$

for any data distribution with well-defined expected loss $\mathcal{R}$. Zeroing the output projection and bias of each appended block makes it the identity, a function-preserving expansion in the sense of network morphisms (Chen et al., 2016). Appendix D gives the statement, its proof, and the cost of one added block.

4 EXPERIMENTS

4.1 SETUP

Source and data. The frozen generator is MiniMax-H3 (MiniMax-AI, 2026) in its reference-to-video path, which accepts one to nine ordered RGB references. We hook twelve capture sites and read the video-stream tokens, as described in Appendix B. The data pipeline draws on RealEstate10K (Zhou et al., 2018), ScanNet++ (Yeshwanth et al., 2023), the DL3DV-Meshed derivative (Ling et al., 2024), and Google Scanned Objects (Downs et al., 2022), covering real-estate video, indoor laser scans, reconstructed scenes, and isolated objects.

Source provenance. The activations behind every measurement here come from a locally converted copy of the published checkpoint whose fused QKV projections were assembled as contiguous thirds instead of by per-head deinterleaving, affecting 156 matrices in 52 attention modules. The frozen source is thus a permuted variant of the published attention, and the absolute scores characterize that variant, not MiniMax-H3's own representations. Every arm reads the same container at the same sites, so the relative attributions below stand as measured. Corrected features are being regenerated and the affected results are retained unchanged.

Architecture of record. Section 3 describes the current architecture, of 11,692,042 parameters. Tables 1–5 were measured on the preceding instance of the same three stages, which held 9,535,893. The two differ inside the Compressor and the Reader, not in their interface: the earlier one used 8×8 windows with eight of 64 DCT modes and one mixed row, sixteen latents per site, one read projection, and no shape descriptor or place code. The tables report the earlier instance unchanged.

Protocol. Surface completion is posed on a given support: birth anchors and their per-anchor areas come from the reference surface of the scene, which fixes the region to be completed and, at the requested density, the size of the point population. Section 3.3 and Appendix C give the anchor jitter and the allocation rule; Appendix C also describes the image-only route, in which MapAnything (Keetha et al., 2026) supplies anchors and the coordinate frame from the source images alone. Unless stated otherwise, we integrate 48 explicit Euler steps at a requested density of $\rho = 9000$ points/m^2 over the complete allocated population, with no point or anchor caps. Each scene is evaluated under two initialization seeds and the two results are averaged; scenes then enter aggregates with equal weight. The unit of comparison is the scene. All scores use raw, unaligned coordinates.

Metrics. Precision P is the fraction of generated points within δ of the reference surface and coverage C the fraction of reference points within δ of the generated set, with $F = 2PC/(P + C)$ and $\delta = \sqrt{3} \cdot 0.02\,\mathrm{m} = 3.46\,\mathrm{cm}$. We report the mean and 95th-percentile generated-to-reference distance, the unsquared symmetric Chamfer distance CD_1, and signed normal accuracy N_{30}, the fraction of points whose normal agrees with its nearest reference normal within 30°. Appendix E gives the definitions.

4.2 SURFACE COMPLETION

Table 1 reports the trained model on a four-scene panel spanning ScanNet++ and RealEstate10K. Two of the scenes appear in the checkpoint's training ledger and two are monitored validation scenes absent from it, and the table keeps them apart. On the validation scenes the model reaches 46.80% precision, 87.99% coverage, and 60.98% F-score at 3.46 cm, with a mean generated-to-reference distance of 4.85 cm. On the training scenes the same numbers are 74.68%, 97.65%, and 83.54%; the strongest of them, RE10K 3390, reaches 93.31% precision and 99.81% coverage at 1.38 cm, which is a fit to a scene the model was trained on, not a generalization result. Coverage stays above 84% on every scene: the fixed population spreads over the whole surface instead of concentrating on the easy parts.

What the Head contributes. The output starts from birth anchors that already sit near the surface, so how much of the final geometry the flow itself produces is worth measuring directly. Table 6 in Appendix G scores the saved initial particles and three reference-derived controls against the final rollout on RE10K 3390, a training scene, under identical scoring. Integration raises precision

Table 1: Surface completion on the four-scene panel. *train* scenes appear in the checkpoint's training ledger; *val* scenes are monitored validation scenes absent from it, but they are not an untouched test set. Two initialization seeds are averaged within each scene and scenes are weighted equally. P, C, F and N_{30} are percentages at $\delta = 3.46$ cm; mean and p95 are generated-to-reference distances in cm. All scores use raw, unaligned coordinates over the complete population.

Scene	Split	P	C	F	Mean	p95	N_{30}
ScanNet++ 00a	train	56.04	95.50	70.63	3.83	10.10	28.17
ScanNet++ 01ce	val	52.96	91.14	66.99	4.14	11.00	34.18
RE10K 3390	train	93.31	99.81	96.45	1.38	3.92	61.72
RE10K 2cd	val	40.65	84.83	54.96	5.55	14.31	12.10
Mean, training scenes		74.68	97.65	83.54	2.61	7.01	44.94
Mean, validation scenes		46.80	87.99	60.98	4.85	12.65	23.14
Mean, all		60.74	92.82	72.26	3.73	9.83	34.04

Table 2: Interventions on Memory at inference, with weights, initial population, seed, density and solver held fixed. Four scenes, two seeds averaged within each scene, scenes weighted equally. Distances are cm.

Memory	P	C	Mean	p95	N_{30}
BRIDGE	60.74	92.82	3.73	9.83	34.04
Constant rows only	56.80	93.79	3.99	10.36	27.31
Fixed mixture	55.19	93.42	4.15	10.79	26.24
Memory zeroed	52.71	93.63	4.47	11.65	26.46
Different-scene Memory	52.36	91.80	4.38	11.61	25.73

from 32.07% to 93.31% and normal accuracy from 6.70% to 61.72%, and cuts CD_1 from 4.38 cm to 1.17 cm. The controls locate the initial state: uniform samples in the reference bounding box reach 4.21% precision, a Gaussian fitted to the reference moments 8.15%, and the reference surface perturbed by 10 cm noise 33.95%, so the initial particles are no better than a coarsely perturbed surface. The Head is also the only stage that produces normals at all: every initial state scores near the 6.7% chance level on N_{30}.

4.3 What the Reader reads

Table 2 holds the weights, the initial population, the seed, the density, and the solver fixed, and intervenes only on Memory at inference. *Constant rows only* keeps the spatially constant component of each window and drops the variation content. *Fixed mixture* replaces the learned coefficients $\alpha_{s,w,j,\cdot}$ by a constant of equal weight. *Memory zeroed* zeros the site latents and summary vectors before the read, leaving the Head active. *Different-scene Memory* supplies the Memory of another scene from the same dataset family with the geometry unchanged.

Each intervention lowers precision and normal accuracy. Another scene's Memory costs 8.38 points of precision and 8.31 of normal accuracy and lowers precision on all four scenes, so the Reader is reading scene-specific content rather than a generic prior. Zeroing Memory costs 8.03 points of precision, which bounds what the Head can produce alone. The two compression interventions separate the ingredients of Equation 3: dropping variation content costs 3.94 points of precision and 6.73 of normal accuracy, keeping it with fixed weights 5.55 and 7.80. The learned mixture therefore contributes more than the mere presence of spatial variation. Coverage is nearly flat across the arms, as expected when the initial population already spans the surface.

Consensus. Under a matched adaptation budget from a common checkpoint, removing the consensus costs 5.81 points of precision and 6.65 of normal accuracy, and one aggregated condition beats exposing all thirteen states to the Head by 1.17 points of precision at a difference of 1,278 parameters (Table 5, Appendix F).

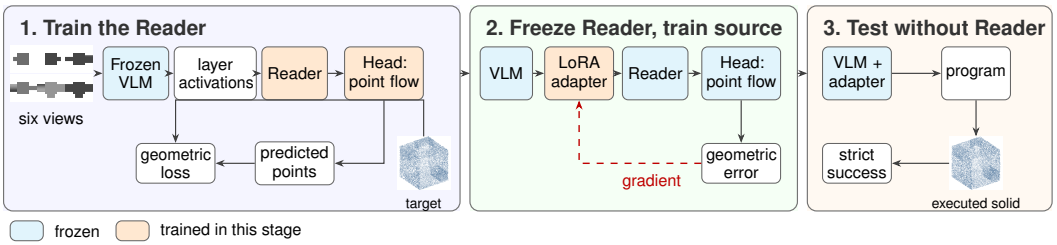


Figure 2: The Reader as a critic for a vision–language model. (1) Reader and Head are trained on the frozen model's activations for its own successful answers. (2) They are frozen, and their geometric error trains a low-rank adapter of the model. (3) The Reader is removed; the adapted model writes a program, the program is executed, and the solid is scored with the unchanged strict criterion. The six views are the task input of one held-out object; the target and executed solids are sampled surfaces of a strict-success answer.

4.4 The Reader as a critic for a vision–language model

The same Reader and Head also run on a different kind of source. We take a frozen vision–language model, Qwen3-VL-4B, show it six views of a solid composed of four parts, and ask for a program that rebuilds the solid; an answer succeeds only if the executed solid reaches volume IoU ≥ 0.90 and surface $F_1 \geq 0.95$. Its activations at 36 layers enter the Reader directly, without a trainable Compressor, and the Head is the oriented-point flow of Section 3.3 with 256 particles, 7.5M parameters in total. We train this Reader on the model's own successful answers, freeze it, and use it as a differentiable critic: its geometric error is backpropagated through the frozen Reader into a new low-rank adapter of the source model. The Reader is then removed, and the adapted model answers on its own (Figure 2).

On 60 held-out objects the model had failed, 10 of which succeed after the adapter path is merged, the adapted model solves 17 to 20, a net repair of $+7$ to $+10$ in every seed, under an unchanged success criterion and without any answer from the Reader at test time. The repair needs a trained critic: a Reader of the same architecture without training changes the count by -1 on average. It does not need the critic to see the particular object: pairing each object with another object's target repairs as much. The trained Reader passes down what it learned about the family of solids, and the source model turns that into better programs for objects it had failed (Appendix I).

Scope of the reported numbers. The absolute scores are modest, above all on validation scenes and for signed normals. The frozen source is the permuted H3 container of Section 4.1, so every table measures a Reader working from rearranged attention; only two validation scenes are available; and signed normals are also scored where Euclidean grouping merges two opposed sheets, a common case (Section 3.2). Corrected features have since been extracted, and the current instance is training on them over 82 scenes; Appendix H reports its latest checkpoint and a schedule issue found in it, and we will report it under the same protocol once training is complete.

Cost. A 48-step rollout is cheap: on one A100 the largest case generates 1,951,983 oriented points in 39.7 s at 2.43 GiB peak memory, with Memory built once and reused across all steps and groups.

5 Conclusion

BRIDGE reads the intermediate representations of a frozen model and turns them into the structure a task requires: a Compressor reduces the activation field to a fixed-size Memory, a Reader lets each output element probe it with its own evolving state, and a Head advances that state. For sparse-view surface completion from a video model, interventions attribute the result to scene-specific Memory, the learned spectral mixture, and the consensus between evidence and geometry. For a vision–language model, the same Reader, frozen, becomes a critic whose geometric error repairs the model's held-out failures after the critic is gone. The source and the Head are interface choices; the Reader is what carries over.

REFERENCES

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via self-distillation with video diffusion models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.19296.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating learning via knowledge transfer. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05641.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A high-quality dataset of 3D scanned household items. *arXiv preprint arXiv:2204.11918*, 2022. URL https://arxiv.org/abs/2204.11918.

Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J. Mitra. Diffusion 3D features (Diff3F): Decorating untextured shapes with distilled semantic features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.00430.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Antoine Guédon, Shu Nakamura, Nicolas Dufour, Jiahui Lei, Ko Nishino, and Angjoo Kanazawa. Surflo: Consistent 3D surface flow model with global state. *arXiv preprint arXiv:2606.13644*, 2026. URL https://arxiv.org/abs/2606.13644.

Zixuan Huang, Justin Johnson, Shoubhik Debnath, James M. Rehg, and Chao-Yuan Wu. PointInfinity: Resolution-invariant point diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10050–10060, 2024. URL https://arxiv.org/abs/2404.03566.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision*, 2026. URL https://arxiv.org/abs/2509.13414.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ling_DL3DV-10K_A_Large-Scale_Scene_Dataset_for_Deep_Learning-based_3D_Vision_CVPR_2024_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. ReconX: Reconstruct any scene from sparse views with video diffusion model. *IEEE Transactions on Image Processing*, 35:2305–2319, 2026. doi: 10.1109/TIP.2026.3666733. URL https://arxiv.org/abs/2408.16767.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.14334.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B. Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.00785.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. PC2: Projection-conditioned point cloud diffusion for single-image 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12923–12932, 2023. URL https://arxiv.org/abs/2302.10668.

MiniMax-AI. MiniMax-H3: Official repository. https://github.com/MiniMax-AI/MiniMax-H3, 2026. Repository revision d21241f0a4b3acbb34c97dae47fa417b7065e438; accessed September 18, 2026.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017. URL https://arxiv.org/abs/1612.00593.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.11651.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.14132.

Chenfeng Xu, Huan Ling, Sanja Fidler, and Or Litany. 3DiffTection: 3D object detection with geometry-aware diffusion features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.01010.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yeshwanth_ScanNet_A_High-Fidelity_Dataset_of_3D_Indoor_Scenes_ICCV_2023_paper.html.

Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. SpatialCrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2505.11992.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018. doi: 10.1145/3197517.3201323. URL https://arxiv.org/abs/1805.09817.

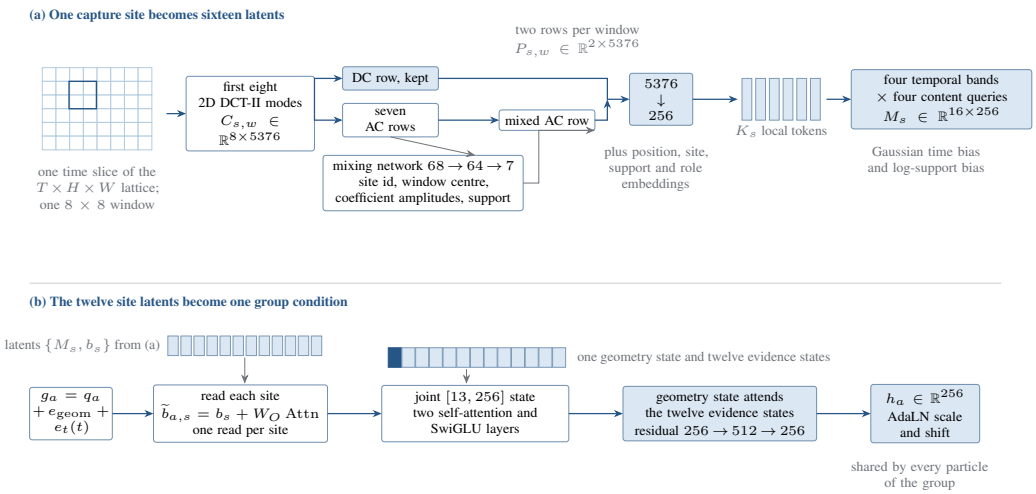


Figure 3: The Compressor and the Reader in detail. (a) The complete DCT basis of each 4×4 window gives one constant row and four adaptively mixed variation rows before the shared wide projection, and sixty-four learned latents resample the resulting local tokens of a site. (b) A probe reads every site through sixteen slots, the joint thirteen-state consensus lets evidence and geometry agree, and the readout conditions every particle of the group.

Table 3: Trainable parameters by stage, from `named_parameters()` of the instantiated network. Fourier frequencies are fixed buffers; the frozen source model, its vision–language encoder, and its autoencoder are external and counted separately.

Stage	Contents	Tensors	Parameters
Compressor	window spectrum, variation mixer, resampler	44	3,903,236
Reader	probe encoder	10	158,848
	slot read and consensus	35	3,547,904
	condition readout	16	527,616
Head	six conditional blocks, two peer exchanges	69	3,554,438
Total		174	11,692,042

A Additional related work

Feed-forward pointmap reconstruction. DUSt3R regresses aligned pointmaps from image pairs without known cameras (Wang et al., 2024), VGGT predicts cameras, depth, and tracks for a whole view set in one pass (Wang et al., 2025), and MapAnything unifies these outputs metrically for a variable number of views (Keetha et al., 2026). All three emit one point per input pixel: output density and extent follow the images, not the surface, and there is no notion of a target population. BRIDGE consumes MapAnything as its image-only initializer and then completes the surface at a requested density the initializer does not determine.

Scaling and capacity. Rectified-flow transformers show favorable scaling behavior in image synthesis (Esser et al., 2024), which motivates keeping the Head's update a plain conditional flow and placing the task-specific structure in the Reader instead.

B Source activations and tensor interfaces

Three time axes. We distinguish native video time τ, the generator's denoising step ℓ, and flow time t. The solver receives $t \in [0, 1)$, the window encoder receives normalized video-window time, and the capture index identifies the denoising iteration. The twelve zero-based (transformer layer, denoising step) capture sites are

$$\begin{array}{cccccc} (39,0) & (19,8) & (9,8) & (35,16) & (44,22) & (35,26) \\ (29,26) & (33,26) & (35,27) & (39,27) & (35,28) & (49,28). \end{array}$$

Generator input and capture. MiniMax-H3 runs its reference-to-video path with image references, accepting one to nine ordered RGB inputs. References are resized toward short edge 2048 at spatial multiples of 32. The visual autoencoder maps images to a latent grid with 24 channels at spatial stride 16, its 48 posterior channels holding mean and log-variance; the implementation samples the posterior, rounds to FP16, applies fixed channelwise normalization, and stores the condition in FP16. The same images condition the Qwen vision–language encoder, whose selected hidden state has width 5120, and text is encoded through the same stream.

The autoencoder supplies a reference latent $[1, 24, 1, h_i, w_i]$ and a target-video latent $[1, 24, 37, h_t, w_t]$, with axes batch, channel, time, height, width. Patches of size $(1, 2, 2)$ group one temporal cell and two cells along each spatial axis into 96-dimensional rows before projection to 5376. The temporal recipe aligns 120 nominal pixel frames to 124 frames, giving 37 latent frames; spatial dimensions follow the aspect ratio, and $[1, 24, 37, 48, 48]$ is the square case, whose post-patch lattice measures $37 \times 24 \times 24$. Target audio contributes 414 rows of width 32 and Qwen a variable number of width-5120 rows. These streams have separate input projections and form one packed sequence inside the transformer.

Each hook stores the complete packed hidden sequence. The Compressor selects the target-video rows, validates a bijection over the declared target lattice, and orders them by time, row, and column. Reference-image, text, and audio rows influence those features through the generator's own attention. The 11.69M parameter count of Table 3 covers the trainable network only; the generator, the vision–language encoder, the autoencoder, and the auxiliary geometry model are frozen external components.

Statistics and boundary handling. Full-field statistics accumulate in FP64. With $L_s = T_s H_s W_s$ native cells and $D = 5376$ channels, flattening F_s to $L_s \times D$ gives

$$\mu_{s,c} = L_s^{-1} \sum_j F_{s,jc}, \quad r_s^2 = (L_s D)^{-1} \sum_{j,c} F_{s,jc}^2, \quad \sigma_s^2 = (L_s D)^{-1} \sum_{j,c} (F_{s,jc} - \mu_{s,c})^2. \tag{11}$$

Normalization uses one centered-RMS scalar shared across channels at each site. The sixteen two-dimensional DCT-II modes are ordered by $(f_y + f_x, \max(f_y, f_x), f_y)$, and boundary windows build their basis and support counts from their own valid cells. For m valid cells the constant row is the normalized sum divided by $\sqrt{m}$. The per-mode amplitude is $e_j = \sqrt{D^{-1} \sum_c C_{jc}^2}$, and the variation mixer consumes $\log(1 + e_j)$ together with the support fraction $m/16$. Constant fields use zero local coefficients and carry their mean and scale through the summary branch, which projects μ_s / r_s concatenated with $\log r_s$, the log ratio $\log(\sigma_s / r_s)$, and a degeneracy flag, a width of $5376 + 3$.

Resampler biases. Let τ_k be the normalized time of local token k, $u_k \in [0, 1]^2$ its normalized window position, and m_k the valid-cell count of its window. With four band centers $c_b = \mathrm{sigmoid}(\hat{c}_b)$ and widths $w_b = \mathrm{softplus}(\hat{w}_b) + 10^{-3}$, and per-latent place centers p_j and widths ς_j formed the same way, the additive bias for latent j is

$$\beta_{jk} = -\frac{1}{2} \left(\frac{\tau_k - c_{\lfloor j/16 \rfloor}}{w_{\lfloor j/16 \rfloor}} \right)^2 - \frac{1}{2} \left\| \frac{u_k - p_j}{\varsigma_j} \right\|_2^2 + \log m_k, \qquad j = 0, \ldots, 63. \tag{12}$$

The sixteen content slots of a band share its band prior and carry independent place priors. Every latent keeps access to all local tokens. The Compressor returns latents $[1, 12, 64, 256]$, a summary state $[1, 12, 256]$, a validity mask, and a log-count bias.

Slot read, consensus, and readout. Equation 6 collects the successive attention operations that produce one condition. Let b_s be the learned summary of site s plus its statistics state and let $\mathsf{K}_s, \mathsf{V}_s$ be its projected keys and values. The learned identity vector e_{geom} marks the probe state, $e_t(t)$ encodes flow time, W_Q emits the sixteen read slots, $W_O^{(g)}$ are the four grouped output projections, and $\mathrm{Attn}(Q, K, V)$ is scaled dot-product attention. The site reads are

$$g_a = q_a + e_{\mathrm{geom}} + e_t(t), \qquad \widetilde{q}_{a,s} = W_Q \, \mathrm{LN}(g_a + b_s), \qquad R_{a,s} = \mathrm{Attn}(\widetilde{q}_{a,s}, \mathsf{K}_s, \mathsf{V}_s) \in \mathbb{R}^{4096}, \tag{13}$$

Table 4: Principal interfaces for one scene. A is the number of current groups in a pack and Q its padded slots per group; the leading 1 is the scene batch dimension; K_s is the local-token count at site s.

Stage	Tensor	Shape
Compressor	Native video-stream field	$[T, H, W, 5376]$
	Window DCT coefficients	$[T, \lceil H/4 \rceil, \lceil W/4 \rceil, 16, 5376]$
	Local projected tokens at site s	$[K_s, 256]$
	Memory: site latents and summaries	$[1, 12, 64, 256]$, $[1, 12, 256]$
Reader	Current group pack and mask	$[1, A, Q, 6]$, $[1, A, Q]$
	Probe input and output	$[1, A, 327]$, $[1, A, 256]$
	Per-site attention keys and values	$[1, 12, 4, 64, 64]$
	Joint evidence–probe state	$[1, A, 13, 256]$
	Condition	$[1, A, 256]$
Head	Persistent particle state	$[N, 6]$
	Fourier input	$[1, A, Q, 72]$
	Peer latents	$[1, A, 128, 256]$
	Output velocity	$[1, A, Q, 6]$

$$\widetilde{b}_{a,s} = b_s + \sum_{g=1}^{4} W_O^{(g)} \left[R_{a,s}\right]_g , \tag{14}$$

where $[\cdot]_g$ selects the g-th block of four slots. The four projections are $1024 \to 256$ and their sum keeps the state at width 256, so the rank limit of the emission rises from 256 to 1024 while the parameter count is unchanged and nothing downstream widens. The joint state $[g_a, \widetilde{b}_{a,1}, \ldots, \widetilde{b}_{a,12}]$ has shape $[13, 256]$. Two four-head self-attention and SwiGLU layers with hidden width 512 update it; the updated probe state then reads the twelve updated evidence states through residual cross-attention and a residual $256 \to 512 \to 256$ SiLU projection, giving $h_a \in \mathbb{R}^{256}$.

C Initialization, training, and numerics

Coordinate frame. Image-only deployment runs MapAnything (Keetha et al., 2026) on the bound source images alone. For its predicted first source-camera pose (R_0, t_0), all predicted pointmaps are transformed as row vectors by $p_{\text{first}} = (p_{\text{world}} - t_0)R_0$, and normals by $n_{\text{first}} = n_{\text{world}}R_0$. Training positions use the same convention with the label-authoritative first camera, so both protocols express geometry in the frame of the first source view. Source activations keep their native lattice coordinates throughout.

Population protocol. Let A_a be the surface area assigned to birth anchor a. The particle count is $q_a = \text{round}_{\text{even}}(\rho A_a)$, computed once in CPU FP64 with ρ drawn log-uniformly from $[1000, 9000]$ points/m^2 and $N = \sum_a q_a$. Training areas derive from conserved reference surface mass, the per-anchor share of area represented by reference samples. Image-only deployment areas come from the valid triangles of the cleaned auxiliary pointmaps and are shared among reachable birth anchors. Allocation happens once, before initialization: no later step reallocates, resamples, reorders, or drops an identity.

Jitter and endpoint coupling. The training layout samples a common amplitude uniformly in $[0, 0.05)$ m and then independent per-anchor offsets uniformly in the corresponding volume ball. The base layout reserves a jitter margin, and the perturbed layout retains complete reference coverage within 0.30 m. The augmentation therefore spans bounded birth-location perturbation together with its local area reassignment. The particle prior adds 0.1 m Gaussian noise around the perturbed anchors.

For XYZ coupling, let c_i be the source birth centers and $\epsilon_i, \eta_i \sim \mathcal{N}(0, 0.1^2 I)$ independent offsets. A log-domain squared-distance Sinkhorn plan with $\varepsilon = 0.05$, 200 iterations, and minibatches of at most 256 is followed by a stable greedy bijection π from target slots to source rows, and the hybrid

source is

$$x_{0,j} = c_{\pi(j)} + \sqrt{1-\beta}\,\epsilon_{\pi(j)} + \sqrt{\beta}\,\eta_{\pi(j)}, \qquad \beta = 0.2. \tag{15}$$

Pairing happens strictly inside one birth domain and never across domains. Each source row keeps its own birth center after pairing. Raw directions are drawn independently with covariance $I/3$, and endpoints use reference unit normals. The target sampler assigns each reference row an area weight w_i and chooses λ so that the capped inclusion probabilities $\min(1, \lambda w_i)$ sum to the requested count, separating certainty rows from randomized systematic sampling.

Updates and precision. A layout uses two independently sampled density and endpoint pairs, each reused for sixteen updates with a fresh time draw and fresh grouping jitter, giving 32 correlated augmentations of the same scene. Every optimizer step sums eight scene losses at weight $1/8$. Two AdamW containers cover the network in three groups:

Container	Group	Parameters	Learning rate
1	Head	3,554,438	1.2×10^{-4}
2	Reader: probe encoder and condition readout	686,464	1.2×10^{-4}
2	Compressor, and Reader's slot read and consensus	7,451,140	3.0×10^{-4}

Both containers step once after the full accumulation, with weight decay 0.01 and Adam epsilon 10^{-8}. The schedule warms up linearly over 400 updates and then follows a cosine decay to a floor of 0.1 of the base rate; the largest gradient norms in every recorded run land in the first few dozen updates, before Adam's second moment has any history to damp them. Gradient norms are clipped at 30 for the Head and 3 for the other two groups, matching measured medians of 6.5–18 and 0.18. Finiteness is checked with a batched infinity norm, which cannot overflow on finite inputs the way an L_2 norm can, and a non-finite update skips the step rather than stepping through it.

Weights, optimizer moments, and persistent state stay in FP32 while CUDA activations may use BF16. Frozen source derivatives may be cached, and Memory is rebuilt for each fresh training graph at the current parameter values: no learned state persists across an optimizer update. Group packing covers all particles, sanitizes padding, and scatters velocities back to the original persistent rows. Final unit decoding requires a direction norm above 10^{-8}.

Separate Fourier bands for position and direction. The Head encodes position and raw direction on separate bands because the two live on different scales. Direction components have standard deviation $1/\sqrt{3}$ and measured median $|d_i| = 0.390$, while positions span whole rooms. Under a shared eight-band encoding at unit period the direction channels reach a median phase of 156.6 rad at the top band, nearly 25 full turns, and 42 of the 48 direction phase channels alias: nearby directions map to distant phases and distant ones can collide. Three bands at period 2 bring the median direction phase to 0.78π and leave no aliased channel. This is a property of the function class, independent of the learned weights.

Opposed sheets under Euclidean grouping. The shape descriptor of Section 3.2 reports $\lambda_{\max}$ and $\|\bar{n}\|$ because grouping is purely Euclidean and therefore merges two surfaces of opposite orientation whenever they pass within one reach of a common anchor. On the measured defect set this affects 34.0% of groups. The sign-sensitive statistic separates them: $\|\bar{n}\|$ has median 0.538 in defective groups against 0.910 in clean ones, and the threshold $\|\bar{n}\| < 0.69$ reaches $F_1 = 0.857$ at 0.861 recall and 0.853 precision, where the sign-blind $\lambda_{\max}$ reaches only 0.721. A linear layer cannot recover a norm from the components of $\bar{n}$, so $\|\bar{n}\|$ is passed explicitly. Giving each side of a thin structure its own anchor does not fix the grouping: measured defect rate rises from 34.0% to 38.6%, because the median spacing between opposed sheets is $2.5\,\mathrm{cm}$ against the roughly $0.28\,\mathrm{m}$ anchor spacing that coverage requires, and the separating plane then cuts each sheet instead of separating the two. The descriptor makes the state visible to the probe; it does not remove it.

D Proofs and cost accounting

Proposition 1 (Function-preserving depth expansion)**.** *Fix the Compressor, state width, grouping rule, peer-exchange locations, and output map. Let $\mathcal{F}_d$ be the family obtained with d conditional*

residual feed-forward blocks, and let $\mathcal{F}_{d+k}$ append k blocks of the same form immediately before the output map. Then $\mathcal{F}_d \subseteq \mathcal{F}_{d+k}$, and Equation 10 holds for any common data distribution with well-defined expected loss $\mathcal{R}(f)$.

Proof. For particle features V and group condition H, a conditional residual block has the form

$$B_\phi(V, H) = V + W_2 \,\mathrm{SiLU}(W_1 \,[\mathrm{LN}(V) \odot (1 + s(H)) + b(H)] + b_1) + b_2, \tag{16}$$

where $s(H)$ and $b(H)$ are the condition-dependent scale and shift. Set $W_2 = 0$ and $b_2 = 0$ in each appended block; on the valid domain of the original network every added block is then the identity for all V, H. Copying the original parameters and keeping the same peer-exchange locations, encoding, readout, and output map, the enlarged network returns the same velocity at every state and time. With the same deterministic grouping, initial state, and Euler steps, induction over steps gives the same complete rollout, and every member of $\mathcal{F}_d$ therefore has a representative in $\mathcal{F}_{d+k}$. Taking the infimum of the same risk over a superset yields Equation 10.

Pack identity. For a fixed scene computation, let $\mathcal{P}_1, \ldots, \mathcal{P}_B$ partition the persistent particle indices with each index appearing exactly once. Writing ℓ_i for the coordinate-weighted squared velocity error,

$$\frac{1}{N}\sum_{i=1}^{N} \ell_i = \sum_{b=1}^{B} \frac{1}{N} \sum_{i \in \mathcal{P}_b} \ell_i, \tag{17}$$

and differentiation gives the same gradient in exact arithmetic when every group is complete and all packs share parameters and context graph. The common denominator N preserves particle weighting across packs; the objective does not depend on how the population is cut into packs.

Parameter closure. Table 3 lists the three stages. One extra width-256 conditional residual block contributes 394,496 parameters and one extra peer exchange 525,824; the trained configuration instantiates six blocks and two peer exchanges. The Compressor's two wide projections, $5376 \to 256$ for the window rows and $5379 \to 256$ for the site summary, account for 2,753,280 of its 3,903,236 parameters, so the stage is dominated by the single unavoidable contact with the source width.

Projection and attention costs. For W_s complete 4×4 windows, projecting native cells to width d has a matrix-product term proportional to $16W_sDd$, while the five-row representation of Equation 3 has term $5W_sDd$, a factor of 3.2. The fixed DCT basis costs order $16W_sRD$ with $R = 16$ and is cacheable. The resampler has attention dimensions $64 \times K_s$ per head, a site read has 16×64 per site and group, the consensus has 13×13 per group, and a peer exchange has $128 \times Q$ and $Q \times 128$ per group. Every operation that depends on the current output state is therefore fixed-size in the source representation: the source width and length enter once, in the Compressor.

E METRIC DEFINITIONS

Let $X = \{x_i\}_{i=1}^{N}$ and $Y = \{y_j\}_{j=1}^{M}$ be the complete generated and reference point sets, with $d_i = \min_j \|x_i - y_j\|_2$ and $e_j = \min_i \|y_j - x_i\|_2$. The unsquared and squared symmetric means are

$$\mathrm{CD}_1 = \tfrac{1}{2}\left(N^{-1}\sum_i d_i + M^{-1}\sum_j e_j\right), \qquad \mathrm{CD}_2 = \tfrac{1}{2}\left(N^{-1}\sum_i d_i^2 + M^{-1}\sum_j e_j^2\right), \tag{18}$$

in meters and meters squared respectively. At threshold δ, precision and coverage are $P_\delta = N^{-1}\sum_i \mathbf{1}[d_i \le \delta]$ and $C_\delta = M^{-1}\sum_j \mathbf{1}[e_j \le \delta]$, with $F_\delta = 2P_\delta C_\delta/(P_\delta + C_\delta)$. We use $\delta = \sqrt{3} \cdot 0.02\,\mathrm{m}$, the diagonal of a $2\,\mathrm{cm}$ voxel, and additionally report mean and 95th-percentile generated-to-reference distances. Signed normal accuracy N_{30} uses nearest-reference dot products, and unoriented scores use their absolute values. Scoring covers the complete reference surface, and reported percentiles are averaged over rollouts. All scores in this paper are raw and unaligned: no rigid, scale, or translation fit is applied before scoring.

Table 5: Consensus depth and condition interface under a matched adaptation budget from a common checkpoint, evaluated on seven scenes with equal scene weight, full population and 48 steps. Every paired update shares endpoints, times, groupings, populations and supervision targets. F is reported at 2 cm and CD_1 in cm.

Condition interface	Cons.	Params	P	C	F	CD_1	N_{30}
Aggregated	2	9.54M	62.98	89.58	51.56	2.67	37.77
Full token set	2	9.53M	61.81	89.19	50.59	2.74	36.55
Aggregated	0	8.22M	57.17	90.08	47.15	2.85	31.12
Full token set	0	8.22M	50.58	90.61	42.03	3.15	21.78

Table 6: Point sets scored against the reference surface on RE10K 3390, averaged over two seeds. The first three rows are reference-derived controls, the fourth is the exact saved initial state. P, C, F, N_{30} are percentages and CD_1 is cm.

Point set	P	C	F	CD_1	N_{30}
Uniform in reference box	4.21	61.15	7.88	30.94	6.69
Gaussian at reference moments	8.15	73.26	14.66	18.35	6.69
Reference surface + 10 cm noise	33.95	99.69	50.65	4.15	6.73
Saved initial particles	32.07	98.43	48.38	4.38	6.70
BRIDGE, 48-step flow	93.31	99.81	96.45	1.17	61.72

F CONSENSUS AND THE CONDITION INTERFACE

Two choices inside the Reader decide how much the output elements can share: the depth of the consensus over the $[13, 256]$ state of Section 3.2, and the form in which the result reaches the Head, either one aggregated condition vector or the full set of thirteen states. We train the four combinations from a common checkpoint under a matched budget, with identical endpoints, times, groupings, populations, and supervision targets in every paired update, and evaluate all at full population with 48 steps. Table 5 reports the seven-scene macro average.

Two consensus layers with one aggregated condition, the configuration BRIDGE uses, is best on precision, F-score, Chamfer distance, and normal accuracy. Removing the consensus costs 5.81 points of precision and 6.65 of normal accuracy with the aggregated condition, and 11.22 and 14.77 with the full token interface: the consensus is what makes the retrieved evidence usable. At equal depth the aggregated condition beats the full token interface by 1.17 points of precision while differing by 1,278 parameters, which places the gain in the aggregation itself.

G INITIALIZATION CALIBRATION

H CURRENT INSTANCE

Table 7 scores the 11.69M-parameter instance of Section 3 at U5632 on the three scenes held out from its 82-scene training set. Against its own initial particles the rollout raises precision by 3 to 8 points on every scene, lowers coverage by 41 to 47 points, and leaves CD_1 worse on all three.

Diagnosis: forgetting under a windowed schedule. This run streams its training set through windows: eight scenes stay resident for 768 consecutive updates and are then replaced, with no replay of earlier windows. Table 8 scores the same checkpoint on training scenes as well, split by whether a scene is resident at U5632 or was trained in an earlier window. We also report the relative error of the predicted velocity at $t = 0$, where 1.0 is the error of predicting zero.

On resident scenes the Head fits the target: F-score rises from 0.44–0.52 to 0.80–0.99, and the velocity error falls to 0.40–0.63. On indoor scenes from earlier windows it returns to the held-out level, with velocity errors of 1.02–1.07; the two earlier objects keep part of their fit. The model therefore learns each window and loses much of it once the window is replaced, which is catastrophic forgetting in the

Table 7: Current instance (11.69M parameters) at U5632 of its first pass over 82 training scenes, on three held-out scenes it was never trained on. *initial* scores the sampled initial particles with no model applied; *48 steps* scores the rollout, averaged over the listed number of initialization seeds. The reference is the density-sampled supervision target in the pair frame, not the complete reference surface of Table 1, so the two tables are not directly comparable. Percentages at $\delta = 3.46$ cm; distances in cm; raw, unaligned coordinates.

Scene	State	Seeds	P	C	F	Mean	CD_1	N_{30}
ScanNet++ 00a	initial	1	28.33	97.81	43.93	7.63	4.73	6.74
	48 steps	2	35.83	55.06	43.40	6.37	5.17	20.94
ScanNet++ 024	initial	1	27.86	91.85	42.75	7.66	4.92	6.70
	48 steps	2	31.03	44.71	36.64	6.84	5.72	9.00
RE10K 33a	initial	1	32.43	88.73	47.50	6.97	4.62	6.76
	48 steps	2	35.91	40.65	38.13	6.27	5.64	10.90

Table 8: Current instance at U5632 on the scenes resident in its training window, on training scenes from earlier windows, and on held-out scenes. F, P and C at $\delta = 3.46$ cm against the density-sampled supervision target, one initialization seed, raw unaligned coordinates. *Initial* is the sampled initial state; e_0 is the relative error of the predicted velocity at $t = 0$, where 1.0 is the error of predicting zero.

Split	Scene	F initial	F	P	C	e_0
resident	GSO Paint Maker	0.516	0.993	1.000	0.987	0.40
	RE10K 97a	0.444	0.799	0.734	0.877	0.63
	ScanNet++ 04c	0.461	0.867	0.796	0.952	0.61
earlier window	GSO My Mood Memo	0.301	0.367	0.299	0.473	0.45
	GSO Racoon	0.391	0.468	0.375	0.622	0.51
	RE10K 3390	0.436	0.404	0.355	0.468	1.07
	ScanNet++ 067	0.391	0.386	0.348	0.434	1.04
	ScanNet++ 0e9	0.425	0.457	0.375	0.586	1.02
held out	RE10K 33a	0.475	0.384	0.360	0.412	1.21
	ScanNet++ 00a	0.439	0.444	0.363	0.572	1.07
	ScanNet++ 024	0.427	0.365	0.310	0.446	1.15

sense of McCloskey & Cohen (1989) and Kirkpatrick et al. (2017). It is a property of the schedule, not of the objective or the architecture: the same loss and network reach $P = 1.00$, $C = 0.99$ on a resident object. The oscillation of the held-out score across windows follows from the same cause, because each window's eight scenes resemble the held-out scenes to a different degree. The planned continuation from this checkpoint keeps the objective and architecture and replaces the schedule: a resident pool of about 24 scenes, eight drawn per update so that every resident scene recurs within a few updates, and one scene replaced per device every 128 updates. We do not use this instance for any claim in Section 4.

I Other readouts of the same architecture

The three stages of Section 3 do not depend on which model supplies the activations or on what the Head produces. The studies below change one or both, keeping the Reader. Section 4.4 reports the vision–language study in the main text; the other two are smaller and preliminary. Figure 6 plots their raw series.

Which source the Reader reads (a, b). The same readout, 804,835 parameters with probe, slot read, consensus, and flow Head, is trained once on frozen H3 activations and once on frozen Wan activations, with identical supervision, pairing, and time draws (Figure 4). Only the source differs. On held-out objects the H3-conditioned readout reconstructs better in 118 of 158 paired evaluations ($p = 3.97 \times 10^{-10}$), and on average the gap widens with training (a). On training objects Wan

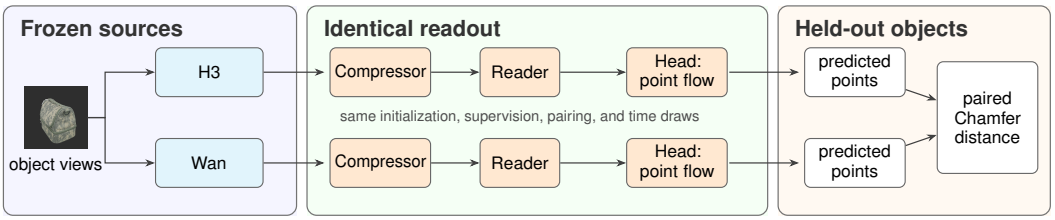


Figure 4: Source comparison. One readout architecture is trained twice from the same initialization, on frozen H3 and on frozen Wan activations of the same objects; supervision, pairing, and time draws are identical, so only the source differs. Held-out objects are scored by paired Chamfer distance.

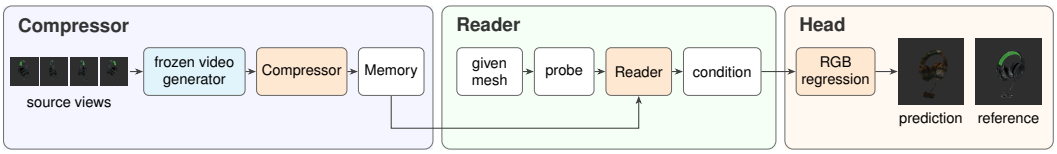


Figure 5: Appearance instance. The Compressor and Reader keep the layout of Section 3; the Head regresses RGB at surface points of the given mesh instead of integrating a flow. The images are Razer's source views and, for a held-out camera of this unseen object, the 14-object model's prediction next to the reference.

fits slightly better, so the advantage is one of generalization, not of fit. The two readouts also use their sources differently: zeroing the condition costs the H3 readout 0.0122 m and the Wan readout 0.0047 m (b), and in the larger 12-training, 4-held-out split the H3 readout pays the larger cost in 24 of 24 pairs. A readout that has learned to use its source generalizes better, and H3 is the source that supports it. One confound works against H3: each toy object has a single anchor whose raw position enters the probe and can stand in for the condition, and coarsening that code widens the gap.

A frozen Reader as a critic for its source (c, d). Section 4.4 describes the protocol. Panel (c) shows the task adaptation on the easy tier, 0 to 313 strict successes of 352; panel (d) shows the net repair on 60 held-out native failures for each arm and seed.

A different Head: surface appearance (e, f). The Head is replaced by a direct regression of RGB at surface points given the object's mesh, with no flow (Figure 5); the network holds 6,647,698 trainable parameters, with a Compressor of the earlier 8×8 shape. On Razer, a headset disjoint from every training object and never seen by the optimizer, held-out-camera PSNR rises from 13.16 to 17.41 dB as the training set grows from 6 to 14 objects, and complete-surface PSNR from 12.81 to 17.40 dB (e); at 30 objects both stay near 17 dB. More updates at a fixed population do not produce this gain: from U4832 to U5184 at 14 objects the held-out score moves by -0.005 dB, and the 6-object run declines from 14.96 dB at U128 to 13.54 dB at U4096. The gain therefore comes from the diversity of the training set. For scale, an earlier two-object instance reaches 29.90 and 37.30 dB on the source views of its own training objects and 17.91 and 19.39 dB on their held-out views, so an object the model has never seen now reaches close to the held-out-view quality of objects it was trained on.

Whether the Reader reads the object in front of it is a separate question, and one donor swap cannot answer it. We therefore score every unseen object's mesh under every unseen object's H3 activations and test, by 100,000 permutations, whether each mesh scores best under its own. With the Compressor and Reader in their current shape (Section 3), two independently trained instances identify their own source above chance, at $p = 0.0024$ over 20 unseen objects and $p = 0.00015$ over 16. Swapping in a donor's activations moves the predicted mean color toward the donor's color, with slope 0.27 ($p = 0.0038$) and 0.30 ($p = 0.00028$), and on the 16-object instance the correlation between predicted and true luminance rises from 0.29 with the source averaged out to 0.63 with the object's own source. The earlier-shape instance used for the PSNR figures shows only a marginal effect ($p = 0.044$), which vanishes on the twelve objects common to all three instances ($p = 0.33$), where the 16-object current-shape instance remains significant ($p = 0.0051$). The readout is coarse: an object's own

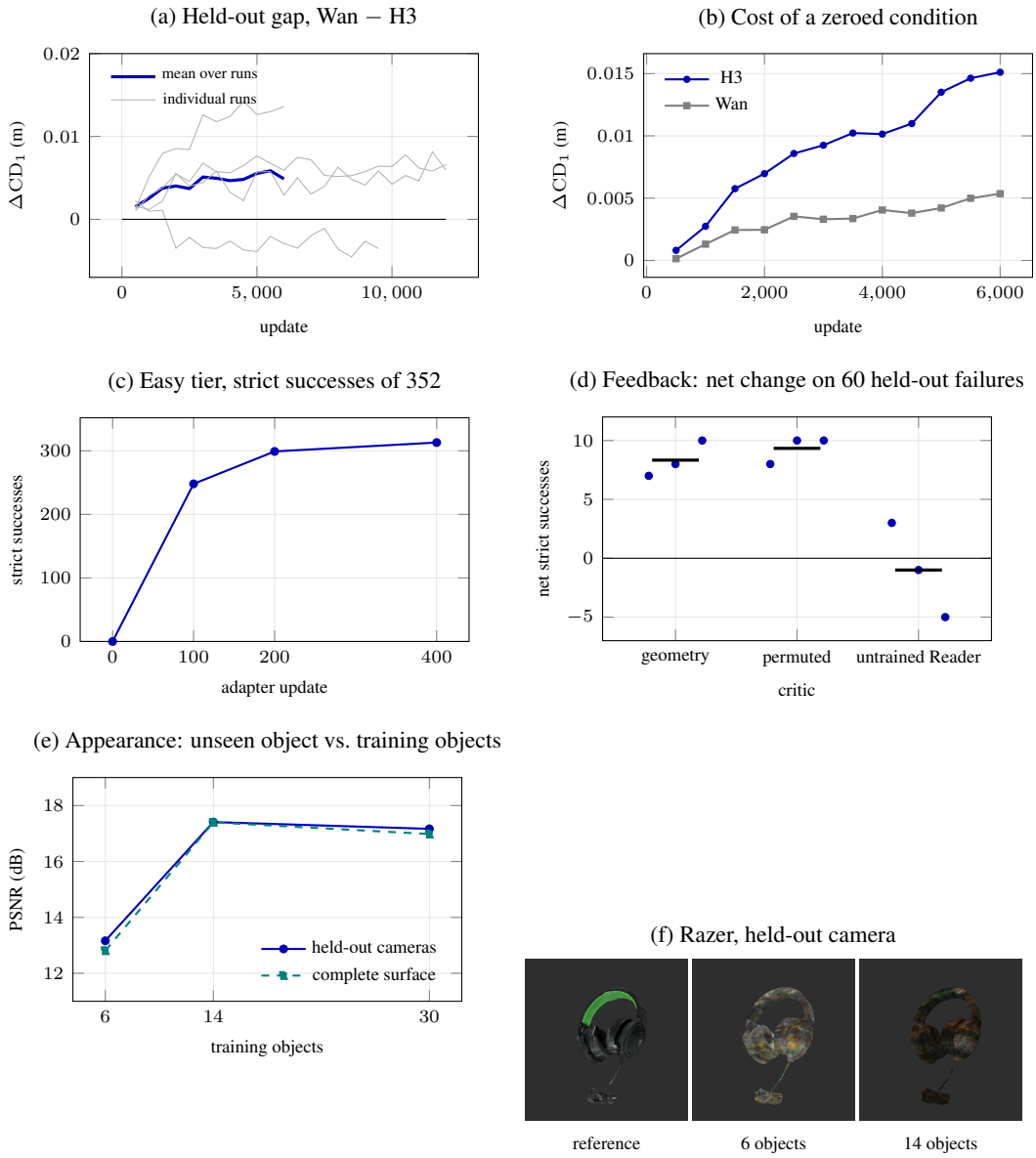
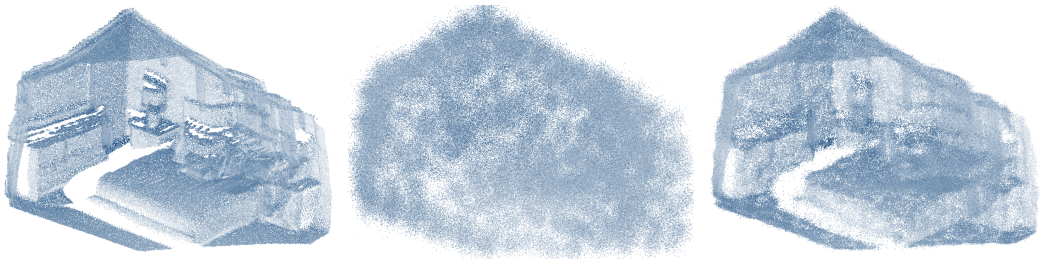


Figure 6: Raw series of the three studies of Appendix I. (a) Held-out CD_1 of the Wan-conditioned readout minus the H3-conditioned one, over two held-out objects; positive favors H3. Thin lines are individual training runs, one of which reverses; the thick line is their mean at the depths all runs reached. (b) Increase in held-out CD_1 when the condition is zeroed, mean over all four seeds and both objects, at the depths every seed reached. (c) Strict successes of the adapted vision–language source on the easy tier. (d) Net change in strict successes on 60 held-out native failures, one dot per seed, bar at the mean. (e) PSNR of the unseen Razer object against training population, on its held-out cameras and its complete surface. (f) One held-out camera of Razer: reference, 6-object model at U4736, 14-object model at U4832.

source ranks first in only 2 of 16 cases, the median advantage over another object's source is 0.83 dB, and cyclically shifting the window positions changes nothing ($|\Delta| \leq 0.021$ dB), so what is read is overall object color and not per-location texture. The current-shape instances are trained from scratch for 4,096 updates and have not yet reached the image quality of the earlier-shape instance, and each configuration has one seed.

J QUALITATIVE RESULT



Reference surface | Initial particles | BRIDGE, 48-step flow

Figure 7: Reference surface, saved initial particles, and the 48-step result of the earlier 9.54M instance on RE10K 3390, a training scene. All panels share one orthographic camera, one light, and one set of bounds, and each draws 328,495 points; every point is shaded by its own normal. Coordinates are raw and unaligned.