ALIGN-VAE: ALIGNING ATTENTION AND GEOMETRY IN VECSET VAES FOR 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Set-based variational autoencoders (VecSet VAEs) represent 3D shapes as compact sets of latent tokens. Although each token originates from a query at a sampled surface location, whether these latents truly align with geometry remains unclear. We investigate this alignment and identify four issues: (i) attention-sink-like tokens that attract queries from across the shape; (ii) spatially periodic repetitions in decoder attention responses; (iii) mismatched normal information in encoder and decoder query features; and (iv) missing supervision of the intrinsic relation between signed distance fields and surface normals. To address these issues, we introduce Align-VAE with four corresponding modifications: (i) a shared bank of learnable decoder tokens provides a workspace for global computation; (ii) distance-biased cross-attention guides retrieval toward nearby predicted spatial anchors; (iii) normal-free encoder queries align the two stages' access to normal information while retaining normals in the encoder keys and values; and (iv) a screened Poisson-inspired regularizer supervises surface values and normal-direction derivatives. These changes align latent retrieval and field prediction with geometry without enlarging the per-shape code. On the ABO dataset, Align-VAE reduces Chamfer distance by 26.6% compared to the standard VecSet VAE.

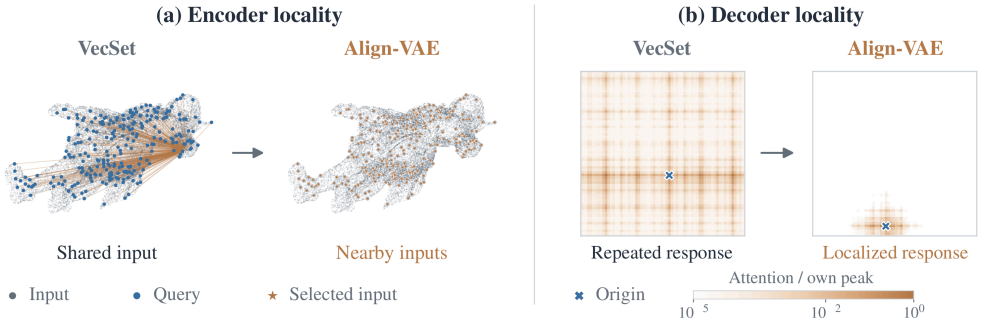


Figure 1: **Improving local responses in VecSet VAEs.** (a) Nearby encoder selections replace a shared input. (b) Decoder responses become localized. The models share an input shape and 8192 stored tokens. Lines show most-attended inputs; crosses mark FPS origins. Each slice shows the median token by surface attention, through its volume peak and normalized by that peak.

1 INTRODUCTION

Compact 3D representations provide an interface between detailed geometry and latent generative models. VecSet variational autoencoders (VecSet VAEs) compress a surface into a fixed number of latent tokens and reconstruct it through an implicit field queried at arbitrary spatial locations (Zhang et al., 2023). Their token budget can be chosen independently of the resolution at which the field is decoded, making them attractive for scalable shape modeling (Zhang et al., 2024; Li et al., 2025a; Team Hunyuan3D et al., 2025). Increasing output sampling density cannot recover detail that the code fails to preserve or the decoder fails to retrieve. Figure 1 previews how Align-VAE improves local responses during encoding and decoding while retaining the stored code budget.

In a point-query VecSet encoder, each latent originates from a query at a sampled surface point. This assigns a spatial origin, but does it also establish a spatial role? Encoder and decoder cross-attention

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.

retrieve features by content similarity, while self-attention mixes information across tokens. None of these operations requires a token to describe its originating neighborhood. Global context helps describe a coherent shape, while local field queries need spatially precise evidence. We hypothesize that sharing limited latent capacity across these roles can leave spatially seeded latents poorly aligned with the local geometry they are meant to represent.

Our profiling of a standard VecSet VAE reveals gaps between spatial origins and geometric roles. Encoder queries often favor shared inputs, while decoder attention concentrates on a small subset of latents (Figures 2(b) and 2(c)). Individual latents can respond at separated locations, and decoded field gradients deviate from surface normals (Figures 3(b) and 3(d)). Encoder queries also use observed normals that arbitrary decoder queries lack. Section 4 explains how these diagnostics motivate changes to local encoding, spatial retrieval, and field supervision.

We introduce Align-VAE to address these issues while retaining the stored VecSet code. A shared bank of learnable decoder tokens supplies a workspace for global computation. Predicted spatial anchors and a distance bias guide retrieval from the remaining shape tokens toward nearby geometry. Normal-free encoder queries remove the normal-information mismatch while keeping observed normals in the support keys and values. Finally, a regularizer inspired by screened Poisson reconstruction (Kazhdan & Hoppe, 2013) constrains surface values and variation along the observed normals. Together, these changes govern how local evidence is encoded, retrieved, and translated into a geometric field with spatial and orientation cues.

Our contributions are threefold:

- **Detailed analysis of VecSet VAEs.** We examine attention allocation, spatial retrieval, query information, and field orientation to characterize the geometric roles of spatially seeded latents throughout encoding and decoding.
- **Targeted improvements.** We introduce Align-VAE, combining a shared decoder bank, distance-biased retrieval, normal-free queries, and normal supervision without enlarging the stored per-shape code.
- **Extensive evaluation.** We evaluate Align-VAE through reconstruction benchmarks, component ablations, and geometric diagnostics. On the ABO dataset, we achieve a 26.6% reduction in Chamfer distance compared to the standard VecSet VAE.

2 Related Work

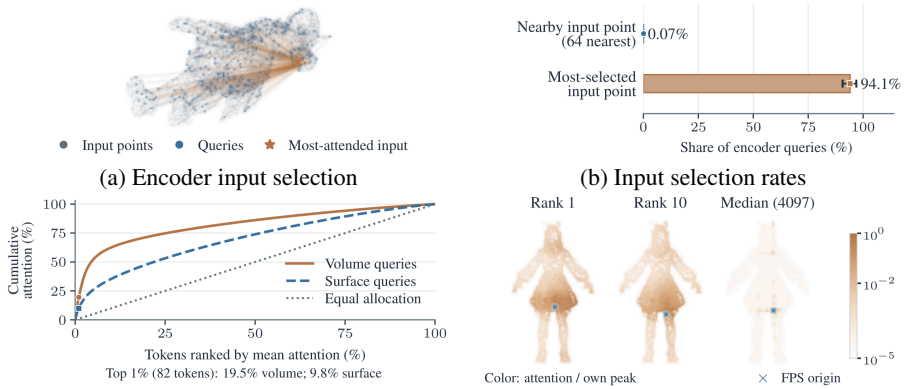
Compact 3D shape latents. 3DILG (Zhang et al., 2022) associates latent features with irregular spatial samples, while 3DShape2VecSet (Zhang et al., 2023) uses attention to connect a compact latent set to continuous field queries. Michelangelo (Zhao et al., 2023) aligns shape latents with image and text features, and CLAY (Zhang et al., 2024), TripoSG (Li et al., 2025a), and HY3D-2.1 (Team Hunyuan3D et al., 2025) scale set-based shape modeling. Dora (Chen et al., 2025) studies sampling and evaluation for 3D VAEs. LATTICE (Lai et al., 2026) anchors latents to a coarse voxel grid; MSVS-VAE (Hao et al., 2026) combines anchored latent hierarchies with local decoding. We analyze the geometric role of VecSet attention.

Spatially structured representations. XCube (Ren et al., 2024) and TRELLIS (Xiang et al., 2025) organize 3D features on sparse spatial grids. Direct3D-S2 and Sparc3D model detailed shapes with sparse representations (Wu et al., 2025; Li et al., 2025b). These methods provide explicit spatial organization; VecSet methods decouple the stored token count from a dense decoding grid. We report accuracy, storage, and decoding cost separately because token widths and layouts differ.

Computational workspace and oriented geometry. Registers provide additional tokens for internal computation in vision transformers (Darcet et al., 2024). We test this workspace interpretation in a 3D VAE by separating an unanchored decoder bank from spatially addressed shape tokens. Poisson and screened Poisson reconstruction connect oriented surface samples to an implicit field (Kazhdan et al., 2006; Kazhdan & Hoppe, 2013). Our regularizer draws on surface interpolation and normal-direction constraints to supervise the learned field during training. Inference evaluates this field without solving a Poisson reconstruction system.

3 Preliminaries

An oriented point set $\mathcal{P} = \{(\mathbf{x}_j, \mathbf{n}_j, s_j)\}_{j=1}^{N}$ contains normalized surface positions, outward unit normals, and scalar sharp-edge labels. Farthest-point sampling (FPS) selects M centers



(a) Encoder input selection

(b) Input selection rates

(c) Cumulative decoder attention

(d) Surface response maps

Figure 2: Attention allocation in a standard VecSet VAE. Local selection uses 64 nearest inputs; the dotted line denotes equal allocation. Surface maps (attention ranks 1, 10, median) use each token's surface peak for log-scale normalization; crosses mark FPS origins. Bars and curves show shape means and 95% intervals; (a, d) share one example.

$\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^{M}$. Each center supplies one query over the N support points. Encoder cross-attention and self-attention yield one latent token of width C_z per center. The autoencoder maps are

$$\boldsymbol{Z} = \mathrm{Enc}(\mathcal{P}, \mathcal{C}) \in \mathbb{R}^{M \times C_z}, \qquad \hat{f}(\mathbf{q}; \boldsymbol{Z}) = \mathrm{Dec}(\mathbf{q}, \boldsymbol{Z}). \tag{1}$$

The decoder refines $\boldsymbol{Z}$ with self-attention, then reads the tokens through cross-attention from an embedded spatial query $\mathbf{q}$. The zero level set of $\hat{f}$ defines the reconstructed surface. The target is a normalized truncated signed distance field (TSDF):

$$y(\mathbf{q}) = r \ \mathrm{clip}(f_{\mathrm{sdf}}(\mathbf{q}), -1/r, 1/r), \tag{2}$$

where $r > 0$ is the inverse truncation half-width and the SDF is positive outside. At a smooth surface point, $y(\mathbf{x}) = 0$ and $\mathbf{n}^\top \nabla y(\mathbf{x}) = r$. The token count M fixes the stored code budget independently of the number of decoder queries.

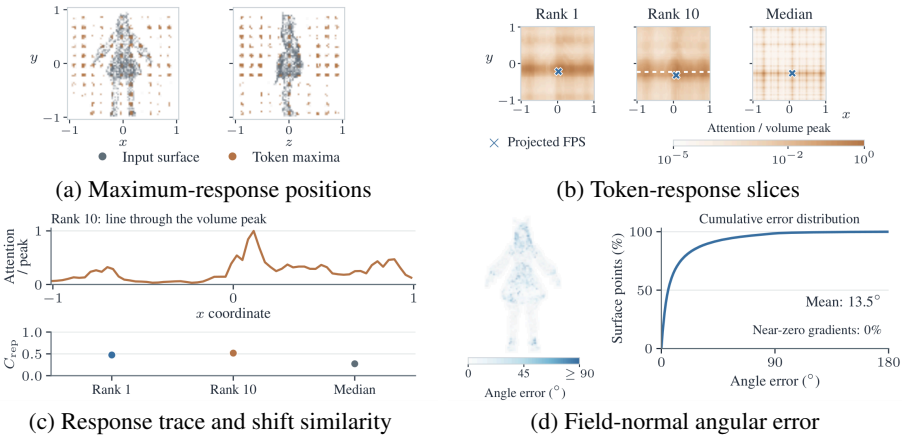
4 Do Spatial Queries Produce Geometrically Aligned Latents?

Figure 2 profiles attention in a standard VecSet VAE. Each encoder query starts at an FPS surface point and gathers information from the input cloud into a latent token. The decoder then reads these tokens from arbitrary 3D query locations to predict an implicit field. Our profiles examine where information is collected, where it is retrieved, and how it determines surface orientation. Appendix B.2.1 illustrates the profiles that motivate the four components of Align-VAE.

4.1 Nonlocal Encoder Selection and Uneven Decoder Attention

Encoder input selection. To test whether distributed query origins yield local reads, we link each of the 8192 FPS encoder queries to the input point with its highest head-averaged attention weight (Figure 2(a)). Figure 2(b) reports the fraction selecting one of its 64 nearest inputs in Euclidean distance and the fraction selecting the shape's most-selected input. The mean rates are 0.07% and 94.1%, respectively: dominant selections largely converge on a shared point despite distributed FPS origins. This behavior resembles attention sinks in LLMs (Xiao et al., 2024) and the use of spatial tokens for global computation in ViTs (Darcet et al., 2024). A per-head analysis also finds two heads with mean local-selection rates above 99% (Appendix B.3), showing that shared selection coexists with local information collection.

Decoder attention concentration. We next measure how field prediction uses the latent budget. A token's *attention mass* is its average attention weight over decoder queries, representing its share of the decoder's attention. We evaluate queries throughout the volume and on the input surface separately. Figure 2(c) ranks tokens by this mass and plots their cumulative share. The top 1% (82 of 8192 tokens) receive 19.5% of volume-query attention and 9.8% of surface-query attention, compared with approximately 1% under equal allocation. This concentration extends the sink-like pattern to latent retrieval and suggests competition between shared computation and local geometric storage. This motivates a shared decoder bank: learnable tokens provide a workspace for global computation, so shape tokens can focus on local geometry without enlarging the stored code.



(a) Maximum-response positions (b) Token-response slices

(c) Response trace and shift similarity (d) Field-normal angular error

Figure 3: Spatial responses and field orientation in standard VecSet. Shape and token ranks match Figure 2(d). Slices are normalized by each token's full-volume peak; crosses mark projected FPS origins. The dashed line gives the trace in (c), alongside shift similarity C_{rep}. Normal maps saturate at 90°. Summaries show shape means and 95% intervals.

4.2 Spatial Repetition in Decoder Cross-Attention

Surface response maps. After measuring how much attention each latent receives, we examine *where* it is read. A token's spatial response is the attention assigned to it as the decoder query moves through 3D space. Figure 2(d) visualizes this response on the surface for tokens ranked first, tenth, and median by surface attention mass. Each map scales colors by that token's own peak response, while crosses mark the originating FPS points. Comparing the colored regions with these origins reveals the spatial extent and displacement of retrieval. These maps show diffuse, smoothly varying responses over broad surface regions, rather than compact peaks around their FPS origins.

Response slices and traces. Figure 3 extends this view into the volume. Figure 3(a) marks the location where each token receives its strongest attention. Figure 3(b) shows the same selected tokens' responses on 2D slices, and Figure 3(c) follows a straight line through a response maximum. The separated peaks show that the same token is retrieved at multiple locations, including regions far from its FPS origin. This is a retrieval-locality mismatch: a token seeded at one surface neighborhood is read as evidence for distant field queries.

Similarity across spatial shifts. To summarize the response structure, we shift each one-dimensional attention curve along its spatial axis and correlate the shifted curve with the original. We denote the highest correlation over a fixed set of nonzero shifts by C_{rep}: larger values indicate more similar responses at separated locations. Mean scores are 0.475, 0.518, and 0.274 for the first, tenth, and median-ranked tokens, respectively (Figure 3(c)). Together with the separated response peaks, this profile motivates an explicit spatial reference for decoder retrieval. We predict an anchor for each shape token and bias attention toward anchors near the query, so retrieval uses both feature similarity and geometric proximity.

4.3 Query-Normal Asymmetry

The encoder–decoder interface introduces a second retrieval mismatch. Encoder queries originate at observed surface points and include their normals. Decoder queries can lie anywhere in space, where a surface normal is unavailable. The encoder can therefore organize information using an orientation cue that the decoder cannot supply when retrieving it. This motivates normal-free encoder queries: removing normals from the query features makes both stages locate evidence without an input normal, while retaining normals in encoder keys and values preserves the observed orientation information. Perturbing only query normals, with input positions and support features fixed, provides a direct test of this dependence through attention and reconstruction (Table 13).

4.4 Field–Normal Misalignment

Geometric alignment also depends on how retrieved features become a field. At the zero level set, the field gradient determines the surface normal, so correct orientation requires the predicted field to vary in the appropriate spatial direction. At observed surface points, we query small offsets

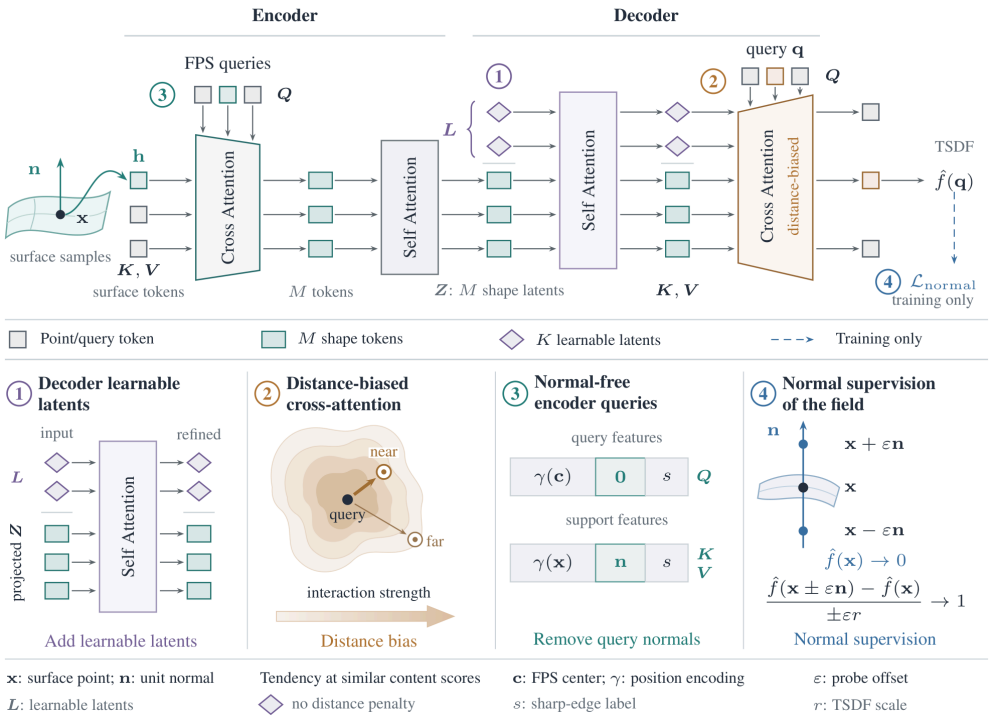


Figure 4: Align-VAE overview. Encoder cross-attention and self-attention produce the M-token code. The decoder prepends K shared learnable tokens before self-attention and retrieves shape tokens with a distance bias. Circled regions show the decoder bank, distance bias, normal-free queries, and normal supervision, respectively. The field receives surface and normal-direction constraints; color bands depict the spatial prior.

along each coordinate axis, estimate the field gradient, and measure its angular difference from the observed outward normal. Figure 3(d) shows a mean error of $13.52°$. This discrepancy exposes an orientation gap in supervision based on field values alone. It motivates coupling field regression with surface and normal-direction constraints: the field should vanish on the observed surface and change along its normals according to the signed-distance target. Our screened Poisson-inspired regularizer imposes these two geometric conditions alongside the attention improvements.

5 Method

Align-VAE retains the M-token variational code and changes how geometric information is collected, retrieved, and supervised. Figure 4 summarizes the architecture. Each component addresses a baseline profile from Section 4; Figure 5 compares the complete models at 8192 latents. Implementation details and training settings appear in Appendix A.

5.1 A Decoder Workspace Alongside Shape Tokens

The decoder mass and spatial profiles in Figures 2(c) and 2(d) motivate a register-like workspace for global computation (Darcet et al., 2024) (Figure 4①). A learned bank $\boldsymbol{L} \in \mathbb{R}^{K \times d}$, shared across shapes, is concatenated with the projected code:

$$\widetilde{\boldsymbol{Z}} = T_{\text{dec}}([\boldsymbol{L} \mid P_{\text{dec}}(\boldsymbol{Z})]). \tag{3}$$

Here P_{dec} projects tokens to width d, $[\cdot \mid \cdot]$ concatenates them, and T_{dec} is the decoder self-attention stack. This interaction makes the bank shape-dependent. Spatial queries read both populations: shape tokens can support local retrieval, while the bank supplies context without anchors or distance penalties. Attention-mass analysis tests this allocation of capacity. The bank adds Kd shared parameters and K decoder tokens, leaving the stored code and KL term unchanged.

Profiling the bank. Figure 5(a) compares the concentration of surface-query attention on shape tokens in VecSet and Align-VAE. Table 12 separates bank share G_{bank} from shape-token read distances and spreads in the cumulative variants.

5.2 NORMAL-FREE ENCODER QUERIES

To remove the query-normal mismatch, we mask normal channels only in the encoder query branch (Figure 4③). With coordinate encoding γ and input projection ϕ_{in}, queries and support features are

$$\mathbf{u}_i = \phi_{\text{in}}([\gamma(\mathbf{c}_i), \mathbf{0}, s_i^{\mathcal{C}}]), \qquad \mathbf{h}_j = \phi_{\text{in}}([\gamma(\mathbf{x}_j), \mathbf{n}_j, s_j]). \tag{4}$$

Encoder cross-attention uses $\mathbf{u}_i$ as queries and $\mathbf{h}_j$ as keys and values. The sharp-edge label is retained: normal-free does not mean identical feature encodings in the two stages. Both stages construct queries without a supplied normal, while the encoder still observes orientation through its support. The locality comparison tests how this change affects information collection.

Profiling normal-free queries. Figure 5(c) compares encoder local selection, the fraction of queries that select a neighboring input point, between the complete models. Table 4 traces this change and local attention mass $P_{\text{enc},k}$ across cumulative variants. Table 13 repeats the perturbations from Section 4.3. Section 6.3 connects these attention changes to reconstruction quality.

5.3 DISTANCE-BIASED RETRIEVAL

Spatially repeated responses motivate an explicit spatial reference for retrieval (Figure 4②). An anchor head predicts $\widehat{\mathbf{a}}_i$ for shape token i, supervised against its FPS center by a regression loss $\mathcal{L}_{\text{anchor}}$. For content logit s_{tn}^h in head h, decoder attention is

$$A_{tn}^h = \text{softmax}_n(s_{tn}^h + b_{tn}), \qquad b_{tn} = \begin{cases} 0, & n \le K, \\ -\dfrac{\beta_\tau}{2\sigma(M)^2}\|\mathbf{q}_t - \text{sg}[\widehat{\mathbf{a}}_{n-K}]\|_2^2, & n > K. \end{cases} \tag{5}$$

Here sg stops the gradient through anchors in the prior; anchor regression still trains them. Nearby anchors receive a smaller penalty, while sufficiently strong content evidence can override the preference. The softmax spans bank and shape tokens together. The strength β_τ is warmed up, and the bandwidth $\sigma(M)$ can adapt to the token budget. Decoding uses predicted anchors and therefore does not require the original FPS coordinates. Attention remains dense; its measured overhead is included in the efficiency comparison.

Profiling distance-biased retrieval. Figure 5(b) compares response correlation C_{rep} for VecSet and Align-VAE. Table 12 reports read distances and spreads for the cumulative variants, using the original FPS centers as reference. A separate bias toggle keeps the latent code and predicted anchors fixed. The bias changes attention but need not remove repetition from raw logits.

5.4 SURFACE AND NORMAL-DIRECTION SUPERVISION

Inspired by screened Poisson reconstruction (Kazhdan & Hoppe, 2013), we supervise both the zero crossing and the field's variation along an observed outward normal (Figure 4④). In the finite-difference formulation, the same latent sample is queried at $\mathbf{x}$ and $\mathbf{x} \pm \varepsilon\mathbf{n}$:

$$d^{\pm} = \frac{\hat{f}(\mathbf{x} \pm \varepsilon\mathbf{n}) - \hat{f}(\mathbf{x})}{\pm\varepsilon r}, \qquad \mathcal{L}_{\text{normal}} = \mathbb{E}\left[\lambda_{\text{surf}}\rho(\hat{f}(\mathbf{x})) + \frac{\lambda_\tau}{2}\sum_{a\in\{+,-\}}\rho(d^a - 1)\right]. \tag{6}$$

The robust penalty ρ is Smooth L1 and $\varepsilon r < 1$ keeps probes in the untruncated band. This constrains the normal direction, not every tangential derivative; full field-normal accuracy is evaluated independently. Appendix A.4 details the three-point sampling and loss reduction.

Profiling normal supervision. Figure 5(d) compares field-normal angular error between standard VecSet and the complete Align-VAE model. The angular error E_{angle} measures agreement between the predicted field gradient and the observed surface normal.

Training objective. We optimize the combined objective:

$$\mathcal{L} = \mathcal{L}_{\text{field}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{normal}}. \tag{7}$$

There is no attention-balancing loss in this four-component method. The cumulative ablation adds the bank (LB), distance bias and anchor regression (DB), normal-free queries (PQ), and normal supervision (NS), in that order.

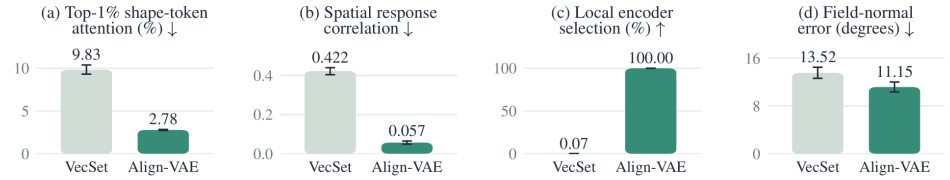


Figure 5: VecSet (gray) and Align-VAE (teal): 8192 shape latents, 64 matched shapes. Bars show means and 95% shape-bootstrap intervals. Panel (a) conditions mass on shape tokens.

Table 1: Median reconstruction scores on Objaverse, ABO, and in-the-wild shapes. Lower CD ($\times 10^3$) and higher F1 (%) are better. Throughout the main text, **bold** and underlined scores mark the best and second-best per column. † marks our reproductions of VecSet and LATTICE.

Method	Tokens	Objaverse			ABO			In-the-wild		
		CD↓	F1@ 0.01↑	F1@ 0.001↑	CD↓	F1@ 0.01↑	F1@ 0.001↑	CD↓	F1@ 0.01↑	F1@ 0.001↑
Dora	4096	5.208	98.90	21.15	5.052	98.49	22.31	4.317	99.97	24.03
HY3D-2.1	4096	6.593	97.72	24.56	6.295	98.28	25.01	11.947	94.92	25.52
VecSet VAE†	4096	4.797	98.97	28.93	5.143	98.74	26.32	3.017	99.83	44.33
VecSet VAE†	8192	3.773	99.62	34.57	3.927	99.54	34.53	2.479	99.95	52.91
LATTICE†	4096	5.529	98.11	26.20	5.490	97.83	24.24	3.773	99.68	34.86
LATTICE†	8192	4.587	99.17	29.12	4.649	98.92	28.87	3.122	99.89	42.39
Align-VAE	4096	3.923	99.57	34.78	4.011	99.68	31.53	2.523	99.96	50.60
Align-VAE	8192	**3.079**	**99.88**	**42.35**	**2.884**	**99.90**	**41.20**	**2.238**	**99.99**	**54.68**

6 Experiments

Section 4 shows that spatially seeded tokens can collect and retrieve information far from their intended neighborhoods. We now ask whether aligning these operations improves reconstruction, which components contribute to the gains, and how their attention changes. Chamfer distance (CD) measures surface error, while F1 measures agreement within a distance threshold; lower CD and higher F1 are better. Appendix B.1 gives their mathematical definitions.

6.1 Reconstruction Quality and Compactness

Better reconstruction from the same code budget. Table 1 compares Align-VAE with standard VecSet VAE, LATTICE, and released 3D autoencoders. At 4096 tokens, Align-VAE lowers CD relative to VecSet by 18.2% on Objaverse (Deitke et al., 2023), 22.0% on ABO (Collins et al., 2022), and 16.4% on in-the-wild shapes. The gain reaches 26.6% on ABO at 8192 tokens. F1@.001 also improves across all three datasets, indicating closer surface agreement. At 8192 tokens, Align-VAE leads every reported metric in Table 1.

Reconstruction across Dora Bench levels. Table 2 extends the comparison to the four levels of Dora Bench (Chen et al., 2025). At 8192 tokens, Align-VAE reduces CD by 48.1% overall relative to VecSet, from 4.804 to 2.495, while F1@.001 increases from 56.87% to 60.22%. The largest CD improvement appears on Level 4, where the error falls from 11.883 to 2.884. Align-VAE achieves the lowest mean CD on all four levels, while VecSet retains the highest F1@.001 on Level 3. At 4096 tokens, overall CD falls from 3.402 to 2.804, indicating that geometric alignment benefits reconstruction at both latent budgets.

Preserving local structure. Figure 6 compares Align-VAE with VecSet and LATTICE at 8192 tokens on selected shapes. The enlarged views show better separated railing bars, clearer screw threads, and raised tire-tread blocks. These local differences are difficult to judge from the overall silhouette alone. These improvements need not reduce whole-shape reconstruction error.

Storage and cost. At 4096 tokens, both models store 512 KiB per shape. The shared bank increases the decoder sequence from M to $M + K$ tokens while preserving the MC_z stored latent values. Latent refinement takes 56.32 ms for Align-VAE and 53.99 ms for VecSet (Table 10).

Table 2: Mean reconstruction scores on Dora Bench. CD is scaled by 10^3; F1 denotes F1@0.001 in percent. Overall averages valid objects across all levels. Method notation follows Table 1.

Method	Tokens	Level 1		Level 2		Level 3		Level 4		Overall	
		CD↓	F1↑	CD↓	F1↑	CD↓	F1↑	CD↓	F1↑	CD↓	F1↑
Dora	4096	3.927	13.57	4.089	21.41	3.932	30.19	4.083	26.25	4.008	22.85
HY3D-2.1	4096	99.117	11.78	66.173	17.04	58.317	15.15	13.855	24.43	59.365	17.10
VecSet VAE†	4096	2.683	69.97	2.132	62.24	3.030	49.40	5.761	33.77	3.402	53.84
VecSet VAE†	8192	2.547	74.21	2.053	64.00	2.732	**52.65**	11.883	36.61	4.804	56.87
LATTICE†	4096	2.990	64.18	2.937	55.14	3.557	46.29	4.792	29.23	3.569	48.71
LATTICE†	8192	3.033	65.97	2.658	56.27	3.320	49.22	4.086	35.73	3.278	51.57
Align-VAE	4096	2.753	71.99	2.144	61.75	3.016	49.31	3.305	40.40	2.804	55.86
Align-VAE	8192	**2.445**	**79.87**	**1.943**	**65.21**	**2.709**	52.01	**2.884**	**43.79**	**2.495**	**60.22**

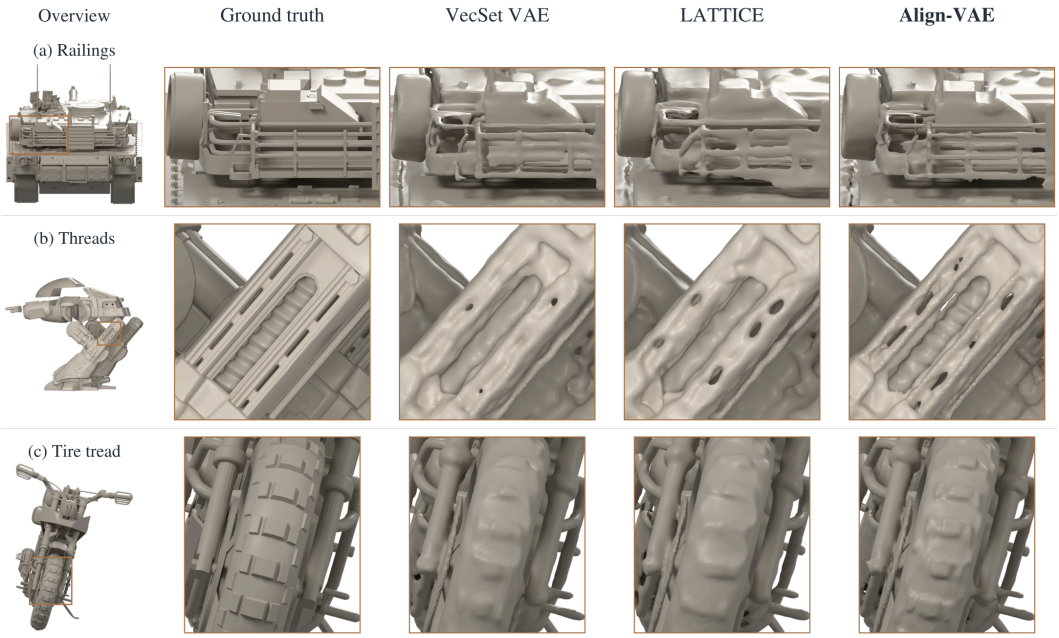


Figure 6: Local geometry at 8192 tokens on selected shapes. Ground-truth overviews mark (a) rear railings, (b) screw threads, and (c) tire tread. Each row uses the same camera, material, lighting, and crop across methods. Details are rendered directly at higher resolution.

6.2 Cumulative Ablation

A0 is standard VecSet. A1 adds a *learnable decoder bank* (LB) for shared context; A2 adds *distance-biased retrieval* (DB) around predicted anchors; A3 uses *normal-free position queries* (PQ); and A4 adds *normal supervision* (NS) through surface and normal-direction constraints. Each variant retains the preceding changes at 4096 tokens.

Table 3 shows a steady reduction in CD from 3.723 to 2.526 (32.2%), while F1@.001 rises from 35.15% to 49.12%. The decoder bank supplies the largest initial gain; distance bias and normal-free queries further improve reconstruction. Sharp-point coverage (S-F) also increases, showing that the gains extend to fine geometry.

6.3 How Does Geometric Alignment Improve?

Figure 5 revisits the profiles from Section 4 at 8192 latents. Relative to VecSet, Align-VAE reduces response correlation from 0.422 to 0.057 and top-1% shape-token attention from 9.83% to 2.78%. Local encoder selection rises from 0.073% to nearly 100%, while normal error falls from 13.52° to 11.15°. Table 4 examines the component contributions through cumulative variants at 4096 tokens, using the same sequence of additions as the reconstruction ablation in Table 3.

Table 3: Cumulative ablation at 4096 tokens; A0–A4 are defined in Section 6.2. CD is scaled by 10^3; F1 and sharp-point coverage S-F are percentages; normal error E_{angle} is in degrees.

Variant	Component				Metric			
	Bank	Distance bias	Normal-free queries	Normal supervision	CD↓	F1@.001↑	S-F↑	E_{angle}↓
A0	×	×	×	×	3.723	35.15	97.30	12.83
A1	✓	×	×	×	2.960	42.70	98.30	11.92
A2	✓	✓	×	×	2.809	45.99	98.44	<u>11.61</u>
A3	✓	✓	✓	×	<u>2.723</u>	<u>46.71</u>	<u>99.11</u>	**10.82**
A4	✓	✓	✓	✓	**2.526**	**49.12**	**99.50**	14.48

Table 4: Geometric behavior of A0–A4, defined in Section 6.2. Local selection measures how often an encoder query selects one of its 64 nearest inputs; local mass sums attention to those neighbors. Lower response correlation and normal error indicate less spatial repetition and better orientation. Bank mass describes attention to shared tokens and has no preferred direction.

Variant	Local selection L_{enc} (%)↑	Local mass $P_{\text{enc},k}$ (%)↑	Bank mass G_{bank} (%)	Response corr. C_{rep}↓	Normal error E_{angle} (deg)↓
A0	<u>0.076</u>	3.05	0.00	0.392	12.83
A1	0.063	13.80	31.19	0.344	11.92
A2	0.043	9.77	54.27	0.052	<u>11.61</u>
A3	**100.000**	**53.82**	53.17	<u>0.034</u>	**10.82**
A4	**100.000**	<u>46.73</u>	48.69	**0.031**	14.48

Separating context from spatial retrieval. In A1, the learnable bank receives 31.19% of surface-query attention, providing a route for shared context. Adding distance bias in A2 reduces the mean read distance from 0.175 to 0.035 and response correlation from 0.344 to 0.052. These changes address the displaced and repeated retrieval patterns observed in the baseline while the bank remains available for shared context.

Collecting evidence near each encoder query. With normal-free queries, A3 selects a neighboring input point for every encoder query, compared with 0.043% in A2. The attention assigned to nearby inputs also rises from 9.77% to 53.82%, so the change extends beyond the single largest attention weight. Reconstruction remains unchanged when query normals are zeroed, perturbed, or flipped (Table 13). Removing the unavailable query cue thus aligns the encoder's information collection with the decoder's inputs.

Reconstruction and orientation are complementary. The combined A4 variant gives the best reconstruction scores, but its field-normal error rises from 10.82° to 14.48° relative to A3. Surface accuracy and gradient orientation therefore capture different aspects of the learned field. This distinction explains why we inspect geometry and attention alongside reconstruction scores; Appendix C expands these comparisons.

7 Conclusion

A token's spatial origin does not guarantee geometrically aligned attention. Align-VAE combines a shared decoder workspace, normal-free encoder queries, spatially biased retrieval, and field supervision while preserving the stored code budget. The resulting representation improves reconstruction and exhibits more local encoder selection and decoder reads. Reconstruction is also insensitive to perturbing encoder query normals. The shared bank keeps per-shape storage fixed but adds decoder computation, so storage efficiency and inference cost must be evaluated separately. The distinction between surface accuracy and field orientation also shows the value of examining how information moves through a 3D representation, beyond measuring its reconstruction error alone. Future work should investigate how to balance surface reconstruction and field orientation, particularly when normal supervision improves one but not the other.

AI USE STATEMENT

We used generative AI tools to polish and rewrite the manuscript text and to help search the literature and format references. We did not use these tools to formulate hypotheses, design experiments, or interpret results. The authors reviewed all AI-assisted text and checked every reference against its original source. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REFERENCES

Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3D shape variational auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16251–16261, 2025.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21126–21136, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Dehao Hao, Kaiyi Zhang, Tanghui Jia, Xiangjun Gao, Dongyu Yan, Weikai Chen, Zeyu Hu, Lingting Zhu, Yingda Yin, Runze Zhang, Li Yuan, Xin Wang, and Long Quan. MSVS-VAE: Multi-scale anchored VecSet for high-fidelity 3D reconstruction. *arXiv preprint arXiv:2607.24436*, 2026. URL https://arxiv.org/abs/2607.24436.

Michael Kazhdan and Hugues Hoppe. Screened Poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3):29:1–29:13, 2013. doi: 10.1145/2487228.2487237. URL https://hhoppe.com/proj/screenedpoisson/.

Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Symposium on Geometry Processing*, pp. 61–70. The Eurographics Association, 2006. doi: 10.2312/SGP/SGP06/061-070.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Haolin Liu, Qingxiang Lin, Jingwei Huang, Chunchao Guo, and Xiangyu Yue. LATTICE: Democratize high-fidelity 3D generation at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19982–19992, 2026. URL https://arxiv.org/abs/2512.03052.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2025a. doi: 10.1109/TPAMI.2025.3633512.

Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3D: Sparse representation and construction for high-resolution 3D shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025b. URL https://arxiv.org/abs/2505.14521.

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. XCube: Large-scale 3D generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4209–4219, 2024.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, Qingxiang Lin, Zeqiang Lai, Xianghui Yang, Huiwen Shi, Zibo Zhao, Bowen Zhang, Hongyu Yan, Lifu Wang, Sicong Liu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Dongyuan Guo, Junlin Yu, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Shida Wei, Chao Zhang, Yonghao Tan, Yifu Sun, Lin Niu, Shirui Huang, Bojian Zheng, Shu Liu, Shilin Chen, Xiang Yuan, Xiaofeng Yang, Kai Liu, Jianchen Zhu, Peng Chen, Tian Liu, Di Wang, Yuhong Liu, Linus, Jie Jiang, Jingwei Huang, and Chunchao Guo. Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material. *arXiv preprint arXiv:2506.15442*, 2025. URL https://arxiv.org/abs/2506.15442.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, and Yao Yao. Direct3D-S2: Gigascale 3D generation made easy with spatial sparse attention. In *Advances in Neural Information Processing Systems*, volume 38, pp. 170778–170804, 2025. doi: 10.52202/085713-5688.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Biao Zhang, Matthias Nießner, and Peter Wonka. 3DILG: Irregular latent grids for 3D generative modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21871–21885, 2022.

Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics*, 42(4):1–16, 2023. doi: 10.1145/3592442.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. CLAY: A controllable large-scale generative model for creating high-quality 3D assets. *ACM Transactions on Graphics*, 43(4):1–20, 2024. doi: 10.1145/3658146.

Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3D shape generation based on shape-image-text aligned latent representation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73969–73982, 2023.

A Implementation and Training Details

This appendix follows the four components in Figure 4, then collects loss reductions and training settings. Architectural operations already defined in the main text are referenced directly.

A.1 Learnable Decoder Bank

Equation 3 prepends K learned tokens of width d to the projected M-token variational code. Joint decoder self-attention makes the bank shape-dependent, and spatial-query cross-attention reads all $K + M$ refined tokens. The bank has no direct access to the input surface and enters only after the variational bottleneck. Its Kd shared parameters add decoder capacity without changing the stored per-shape code or the KL penalty.

Bank tokens have no assigned coordinates and receive no distance penalty. They provide an unanchored route for context, while the shape tokens support spatially addressed retrieval. Both populations compete in the same softmax, with no prescribed attention quota for the bank.

A.2 Distance-Biased Retrieval

An anchor head predicts $\widehat{\mathbf{a}}_i = g_a(\widetilde{\mathbf{z}}_{K+i})$ from refined shape-token features. Training targets the FPS centers $\mathbf{c}_i$, with the loss averaged over all token coordinates:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{3M} \sum_{i=1}^{M} \sum_{c=1}^{3} h_{\delta_a}(\widehat{a}_{ic} - c_{ic}).$$

Here h_{δ_a} is Huber loss: quadratic $u^2/2$ for residuals $|u| \le \delta_a$, and linear $\delta_a(|u| - \delta_a/2)$ otherwise. The cumulative variants use squared error in place of Huber loss; Table 5 records the applicable settings. Anchors are detached only in the distance prior of equation 5; regression still trains the anchor head and its input features. Decoding therefore needs no original FPS points.

The prior is shared across attention heads. After a content-only warm-up, its strength β_τ rises linearly to β. The bandwidth adapts to the number of shape tokens:

$$\sigma(M) = \sigma_{\text{ref}} \sqrt{M_{\text{ref}}/M}.$$

The reference bandwidth σ_{ref} is specified at token count M_{ref}. This scaling follows sample spacing on a surface; it does not enforce a fixed number of attended neighbors. The biased attention aggregates values inside the residual cross-attention block, followed by normalization and a linear scalar head that predicts the field value $\hat{f}(\mathbf{q}; \boldsymbol{Z})$.

A.3 Normal-Free Encoder Queries

Equation 4 masks only the query normal channels. Support keys and values retain observed normals, and both branches retain their sharp-edge labels. Masking preserves the feature width and introduces no parameters. The encoder and decoder thus construct queries without supplied normals, although their projections and remaining features differ.

Encoder cross-attention uses scaled dot products, query/key normalization, and a softmax over support points. The residual block includes value aggregation and a feed-forward sublayer. Encoder self-attention then refines the M tokens, and a projection predicts the mean $\boldsymbol{\mu}$ and log variance of a diagonal Gaussian posterior (Kingma & Welling, 2014). We sample $\boldsymbol{Z}$ from this posterior during training; the decoder bank is added afterward.

A.4 Surface and Normal-Direction Supervision

For each of N_s surface samples, normalize the outward normal $\mathbf{n}_k$ and evaluate the field at $\mathbf{x}_k$ and $\mathbf{x}_k \pm \varepsilon \mathbf{n}_k$, as in Figure 4. All three queries reuse the same sampled latent set and refined decoder features. The two slopes d_k^+ and d_k^- use the definition in equation 6, with the TSDF scale r in their denominator. Both approximate the field derivative along the normal divided by r, whose target is one at a smooth surface point.

The sample reduction of this three-point constraint is

$$\mathcal{L}_{\text{normal}} = \frac{\lambda_{\text{surf}}}{N_s} \sum_{k=1}^{N_s} \rho(\hat{f}(\mathbf{x}_k)) + \frac{\lambda_\tau}{2N_s} \sum_{k=1}^{N_s} \big[\rho(d_k^+ - 1) + \rho(d_k^- - 1)\big].$$

The Smooth L1 penalty ρ is quadratic $u^2/2$ for $|u| < 1$ and linear $|u| - 1/2$ otherwise. The surface weight is constant; the slope weight λ_τ rises linearly from zero to λ_{normal} during warm-up. The condition $\varepsilon r < 1$ keeps the target probes within the untruncated band; the offset must also be small relative to local geometric variation. The prior is evaluated at each displaced query, so the differences include its effect on the predicted field. These probes constrain variation along the normal. The six coordinate-axis probes in Appendix B.2.1 independently evaluate the complete gradient direction.

A.5 Loss Reductions and Training Settings

The field scale r converts normalized signed distance into the TSDF target y_t defined in Section 3. We average squared errors separately over the volumetric query set $\mathcal{V}$ and near-surface set $\mathcal{N}$, of sizes N_{vol} and N_{near}:

$$\mathcal{L}_{\text{field}} = \frac{1}{N_{\text{vol}}} \sum_{t \in \mathcal{V}} (\hat{f}(\mathbf{q}_t) - y_t)^2 + \frac{\lambda_{\text{near}}}{N_{\text{near}}} \sum_{t \in \mathcal{N}} (\hat{f}(\mathbf{q}_t) - y_t)^2.$$

The standard diagonal-Gaussian KL penalty against a unit Gaussian is averaged over the M posterior tokens and C_z channels, excluding the shared bank. The combined objective in equation 7 is then averaged over shapes. Anchor regression trains the retrieval geometry, while normal supervision acts on the field; neither balances attention mass.

Table 5 collects settings from the frozen training snapshots for the in-the-wild experiments. A0–A4 denote the cumulative variants in Appendix C; "Full" denotes the full-scale Align-VAE reproduction evaluated at 4096 and 8192 tokens. A0–A3 use 100k-step checkpoints and A4 uses 110k steps, so this series is not matched in training duration. Each parameter row states its experimental scope.

Table 5: Training settings for the in-the-wild experiments. Offsets and bandwidths use normalized coordinates; schedule lengths are optimizer steps. Settings are scoped to the listed experiments.

Parameter	Meaning	Setting	Experiments
K	Decoder bank size	256	A1–A4, Full
σ_{ref}	Reference bandwidth	0.097	A2–A4, Full
M_{ref}	Reference token count	4096; 8192	A2–A4; Full
β	Final prior strength	1	A2–A4, Full
Prior schedule	Delay, then ramp	10k, then 5k	A2–A4, Full
λ_{anchor}	Anchor loss weight	0.1	A2–A4, Full
Anchor loss	Coordinate penalty	MSE; Huber ($\delta_a = 0.04$)	A2–A4; Full
N_{vol}, N_{near}	Queries per group	4096; 8192	A0–A4; Full
λ_{near}	Near-surface weight	0.1	A0–A4, Full
λ_{KL}	KL weight	10^{-5}	A0–A4, Full
r	Scale at $M \geq 4096$	128	A0–A4, Full
N_s, ε	Samples, offset	2048, 0.004	A4, Full
λ_{surf}, λ_{normal}	Surface, slope weights	1, 0.05	A4, Full
Slope ramp	Linear warm-up	10k	A4, Full

B Evaluation and Profiling Protocols

B.1 Reconstruction Metrics

Chamfer distance (CD). Let $\mathcal{M}$ and $\widehat{\mathcal{M}}$ denote the ground-truth and reconstructed meshes in normalized coordinates, with surface samples G and P of sizes N_G and N_P. The point-to-mesh

distance is $d(\mathbf{x}, \mathcal{S}) = \min_{\mathbf{y}\in\mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2$; the closest point can lie inside a triangle, on an edge, or at a vertex. CD sums both directional means:

$$\mathrm{CD} = \frac{1}{N_P} \sum_{\mathbf{p}\in P} d(\mathbf{p}, \mathcal{M}) + \frac{1}{N_G} \sum_{\mathbf{g}\in G} d(\mathbf{g}, \widehat{\mathcal{M}}).$$

Distances are unsquared, and the two terms are added without a factor of $1/2$. CD therefore retains the units of the normalized geometry.

F1 score. At threshold δ, precision Prec_δ is the fraction of predicted samples within distance δ of the true mesh; recall Rec_δ is the fraction of true samples within that distance of the reconstructed mesh. Their harmonic mean is

$$\mathrm{F1}_\delta = \frac{2\,\mathrm{Prec}_\delta\,\mathrm{Rec}_\delta}{\mathrm{Prec}_\delta + \mathrm{Rec}_\delta}.$$

We set F1 to zero when both precision and recall are zero. F1@.01 and F1@.001 use thresholds 0.01 and 0.001 in normalized coordinates; a column labeled F uses 0.01. Each score is computed per shape before aggregation across the evaluated objects.

Sharp-region distance and coverage. From a fixed ground-truth sharp set G_{sharp} shared by all methods, we measure one-way distance and coverage:

$$\text{S-CD} = \frac{1}{|G_{\mathrm{sharp}}|} \sum_{\mathbf{g}\in G_{\mathrm{sharp}}} d(\mathbf{g}, \widehat{\mathcal{M}}),$$

$$\text{S-F}_\delta = \frac{1}{|G_{\mathrm{sharp}}|} \sum_{\mathbf{g}\in G_{\mathrm{sharp}}} \mathbf{1}[d(\mathbf{g}, \widehat{\mathcal{M}}) \le \delta].$$

S-F is the fraction of ground-truth sharp points within $\delta = 0.01$ of the reconstruction, unless another threshold is stated. It is one-way coverage and does not include a precision term.

Aggregation and units. Reconstruction tables report medians across shapes, except Dora Bench (Table 2), which reports arithmetic means. CD and S-CD are multiplied by 10^3; F1 and S-F are percentages. Attention and field-normal profiles report means across shapes, as defined below.

B.2 Diagnostic Profiles

B.2.1 Reading and Computing the Profiles

Figure 7 links each diagnostic to its computation. An encoder attention row corresponds to an FPS query and its columns to input points; a decoder row corresponds to a field query and its columns to latent tokens. Each softmax row sums to one. We average weights over heads before computing the main attention statistics, then compute each statistic per shape and average across shapes. Figure 5 compares complete models at 8192 shape tokens using the same inputs, FPS centers, and spatial probes. Table 6 provides a symbol and unit lookup.

(a) Local and repeated encoder selection. For each FPS query, select the input with the largest head-averaged attention weight. Local selection L_{enc} is the fraction of queries whose selected input belongs to their $k = 64$ nearest support points. The most-selected input share is the largest selection count divided by the number of queries. These events may overlap: the two bars in Figure 2(b) are not a partition. The per-head control instead takes each head's argmax first and averages the fraction of local selections across heads.

Local mass $P_{\mathrm{enc},k}$ sums the attention on the same 64 neighbors, then averages over queries. Comparing this mass with local selection separates total local evidence from the strongest link. The decoder bank contributes no encoder rows or support points.

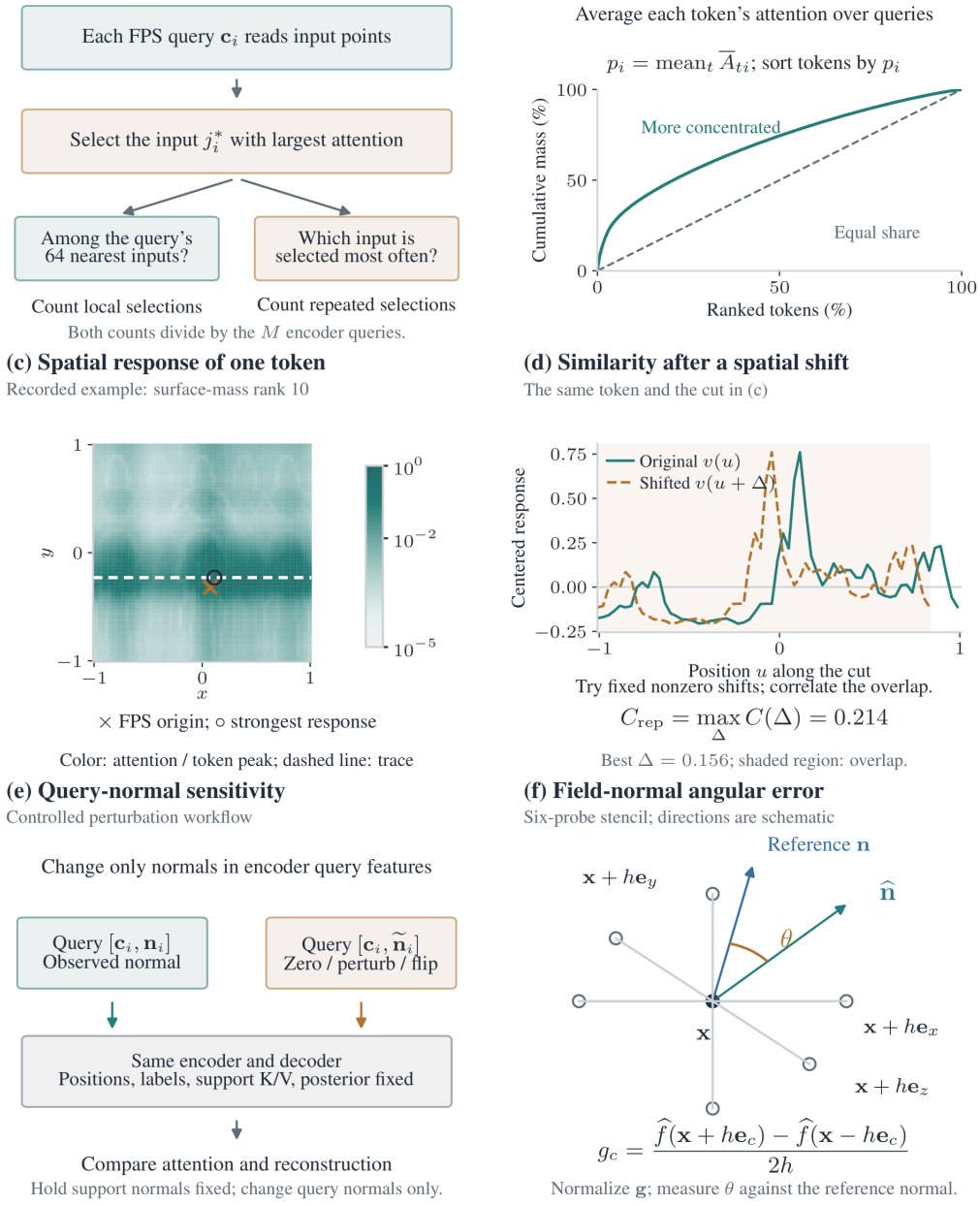


Figure 7: A visual guide to the profiling measurements. (a) Follow each encoder query's strongest attention link, then count local or repeated selections. (b) Average decoder attention over queries, rank tokens, and accumulate their shares. (c) Hold a token fixed and view its response over spatial locations. (d) Shift a response curve and compare the overlapping values; retain the largest correlation over the specified shifts. (e) Change query normals while fixing the other encoder inputs. (f) Estimate the field gradient with six offset queries and compare its direction with the reference normal. Panels (a,e,f) are schematics. Panels (b–d) reuse the first example in Figures 2 and 3; (c,d) use its surface-mass rank-10 token. Panel (d) reports the score for this single trace.

(b) Decoder mass and concentration. Let $\overline{A}_{t,K+i}$ be the head-averaged decoder attention from query t to shape token i, after the K bank columns. Summing over queries and normalizing by total shape-token attention gives its share p_i and the normalized entropy H_{shape}:

$$p_i = \frac{\sum_t \overline{A}_{t,K+i}}{\sum_t \sum_{j=1}^{M} \overline{A}_{t,K+j}}, \qquad H_{\text{shape}} = -\frac{1}{\log M} \sum_{i:p_i>0} p_i \log p_i.$$

Equal allocation gives entropy one; concentration on one token gives zero. Sorting p_i and accumulating it yields the mass curve. Its first $\lceil 0.01M \rceil$ entries sum to the top-1% share $T_{1\%}$. Ranks are computed independently for each model and probe population, excluding the bank. Bank share G_{bank} is the sum over bank columns, averaged over queries; it is zero when $K = 0$. Conditioning on shape tokens uses the denominator above, rather than averaging separately normalized query rows. A zero denominator leaves the conditional profile undefined.

(c) Spatial response, read distance, and spread. Holding a token fixed gives its attention response over query locations. Maps divide each response by its own peak to show spatial structure; color thus indicates relative response, not absolute attention mass. The token's FPS center supplies a common reference for both models. The conditional mean read distance is

$$R_{\text{shape}} = \frac{\sum_t \sum_{i=1}^{M} \overline{A}_{t,K+i} \|\mathbf{q}_t - \mathbf{c}_i\|_2}{\sum_t \sum_{i=1}^{M} \overline{A}_{t,K+i}}.$$

The denominator conditions on shape-token reads, so diverting attention to the bank alone cannot reduce this distance. Restricting both sums to the most-read 1% of shape tokens gives R_{top}. Anchor error, the mean Euclidean distance between predicted anchors and FPS centers, checks the predictor separately from these read distances.

For token i, normalize its response across queries as $w_{ti} = \overline{A}_{t,K+i} / \sum_u \overline{A}_{u,K+i}$. Its response centroid $\bar{\mathbf{q}}_i$ and root-mean-square spread S_i are

$$\bar{\mathbf{q}}_i = \sum_t w_{ti} \mathbf{q}_t, \qquad S_i = \left(\sum_t w_{ti} \|\mathbf{q}_t - \bar{\mathbf{q}}_i\|_2^2 \right)^{1/2}.$$

The mean spread at mass ranks 1, 10, and the median is S_{sel}; S_{top} is the attention-mass-weighted mean of S_i over the top 1%. Zero-mass tokens have undefined conditional statistics. Table 7 includes uniform references that give equal weight to each spatial query. A compact but displaced response can have a large read distance; a broad response has large spread around its own centroid.

(d) Similarity across spatial shifts. For a token's line response $a(u)$, subtract the mean of the entire line to obtain $v(u) = a(u) - \overline{a}$. For shift ℓ, let I_ℓ contain only samples that overlap with the shifted line, without wrapping at the boundaries. The normalized correlation and its maximum are

$$C(\ell) = \frac{\sum_{u \in I_\ell} v(u) v(u+\ell)}{\sqrt{\sum_{u \in I_\ell} v(u)^2 \sum_{u \in I_\ell} v(u+\ell)^2}}, \qquad C_{\text{rep}} = \max_{\ell \in \Lambda} C(\ell).$$

The fixed set Λ contains nonzero spatial shifts with at least half of the line overlapping. Constant traces and zero denominators have undefined correlation. Values near one indicate similar responses at separated locations. The score is read alongside the response peaks; it alone does not establish a common spatial period. Scores are averaged over the three coordinate directions within each shape. Figure 3 and Table 8 keep surface-mass ranks 1, 10, and the median separate; Figure 5 also averages these ranks for its model comparison.

(e) Query-normal sensitivity. Keep positions, labels, support features, and the posterior sampling rule fixed. Change only query normals: retain them, set them to zero, perturb and renormalize them, or randomly flip their signs. Table 13 uses Gaussian perturbations with standard deviation 0.2 and flip probability 0.5. Local selection, local mass, and reconstruction CD measure dependence on this query cue. Normal-free queries ignore these changes while support normals remain available; the table reports this controlled comparison for A0, A2, and A3.

(f) Field orientation. Six queries at positive and negative offsets along the three coordinate axes estimate the gradient at surface point $\mathbf{x}_k$:

$$[\mathbf{g}_k]_c = \frac{\hat{f}(\mathbf{x}_k + h\mathbf{e}_c) - \hat{f}(\mathbf{x}_k - h\mathbf{e}_c)}{2h}, \qquad c = 1, 2, 3,$$

where $\mathbf{e}_c$ is a coordinate unit vector. The evaluation step h is independent of the training offset ε; the six probes evaluate the complete gradient direction, while the three training probes constrain variation along the reference normal. For a fixed threshold $\eta > 0$, the valid index set is $\mathcal{I}_{\text{valid}} = \{k : \|\mathbf{g}_k\|_2 > \eta\}$. Valid gradients yield unit vectors $\widehat{\mathbf{n}}_k = \mathbf{g}_k / \|\mathbf{g}_k\|_2$. Their angular error against unit reference normals $\mathbf{n}_k$ is

$$E_{\text{angle}} = \frac{180}{\pi |\mathcal{I}_{\text{valid}}|} \sum_{k \in \mathcal{I}_{\text{valid}}} \arccos\big(\text{clip}(\widehat{\mathbf{n}}_k^\top \mathbf{n}_k, -1, 1)\big) .$$

Lower error means closer agreement in surface orientation. For N_s evaluation points, the near-zero-gradient rate is $1 - |\mathcal{I}_{\text{valid}}|/N_s$; an empty valid set has undefined angular error. Paired angular-error differences use common valid points and retain each model's invalid rate. Compared models share surface samples and subsets. The in-the-wild profiles use 8192 points, $h = 0.002$, and $\eta = 10^{-6}$.

Table 6: Lookup for diagnostic names and symbols used in the figures and tables. Coordinate units refer to the normalized geometry. Fractions are displayed as percentages where indicated.

Metric	Symbol	Meaning	Unit
Local selection	L_{enc}	Queries selecting one of 64 neighbors	Fraction
Local mass	$P_{\text{enc},k}$	Attention on 64 nearest inputs	Fraction
Most-selected input	—	Largest share of queries selecting one input	Fraction
Token mass	p_i	Conditional share of a shape token	Fraction
Top-1% share	$T_{1\%}$	Mass on the top 1% of shape tokens	Fraction
Mass entropy	H_{shape}	Evenness of shape-token use	Unitless
Bank share	G_{bank}	Mass on the decoder bank	Fraction
Read distance	R_{shape}, R_{top}	Distance from query to token's FPS center	Coordinate
Response spread	S_i, S_{sel}, S_{top}	Extent around response centroid	Coordinate
Shift correlation	C_{rep}	Largest nonzero-shift correlation	Unitless
Normal error	E_{angle}	Gradient angle to surface normal	Degrees

Table 7: Baseline attention diagnostics: means [95% shape-bootstrap intervals]. Distances and spreads use canonical coordinates; lower values indicate greater spatial locality.

Statistic	Volume queries	Input-surface queries
$100T_{1\%}$	19.50 [18.48, 20.61]	9.83 [9.32, 10.39]
H_{shape}	0.875 [0.871, 0.878]	0.960 [0.959, 0.962]
$R_{\text{shape}}\downarrow$	0.953 [0.937, 0.969]	0.304 [0.288, 0.319]
$R_{\text{top}}\downarrow$	1.030 [1.001, 1.061]	0.393 [0.370, 0.418]
Uniform reference for R_{top}	1.244 [1.224, 1.266]	0.764 [0.726, 0.804]
$S_{\text{top}}\downarrow$ (mass weighted)	0.893 [0.882, 0.905]	0.354 [0.338, 0.371]
Uniform-query spread	1.029 [1.029, 1.029]	0.619 [0.593, 0.645]

Table 8: Spatial-shift response correlation in the baseline, with 95% confidence intervals. Tokens are ranked by surface attention mass; lines pass through either the FPS origin or the strongest response. Lower correlation indicates less spatial repetition; bold and underlined estimates mark the lowest and second-lowest per column.

Surface mass rank	Lines through FPS origin $\downarrow$	Lines through volume peak $\downarrow$
1	<u>0.493</u> [0.456, 0.531]	<u>0.475</u> [0.434, 0.515]
10	0.548 [0.524, 0.575]	0.518 [0.491, 0.547]
4097 (median)	**0.292** [0.268, 0.315]	**0.274** [0.250, 0.296]

B.3 Additional Baseline Attention Measurements

Shared selection can coexist with local attention. The head-averaged encoder selects a nearby input for only 0.073% of queries, yet two individual heads have local-selection rates above 99%. Local attention mass is 7.22%, while the most-selected input receives the dominant selection for 94.06% of queries. These measures capture different aspects of the same attention pattern: a common strongest selection can coexist with useful local evidence in other heads.

Attention mass and spatial extent answer different questions. Table 7 compares decoder reads near the surface and throughout the volume. The top 1% of tokens receive 9.83% of surface attention and 19.50% of volume attention. Surface responses are more compact, but their mass remains concentrated. In the spatial maps, each token's response is divided by its own peak; colors therefore show where a token is read, rather than its absolute share of attention. The FPS markers provide the geometric reference for reading those maps.

B.4 Spatial Responses and Field Orientation

Peak locations and response curves. Figure 3 follows tokens ranked first, tenth, and median by surface attention. A maximum-response position shows where a token is read most strongly; slices and line traces reveal whether it is also retrieved at other locations. Table 8 reports spatial-shift correlations for lines through either the token's FPS origin or its response peak. Positive correlations in both views support spatially extended token responses.

Field orientation. The baseline's mean angular error is 13.52°, while its mean per-shape median and 90th percentile are 5.64° and 37.12°. The broad error distribution shows why accurate field values alone do not fully characterize the reconstructed geometry. The angular profile complements surface distance by measuring the direction of local field variation.

C Additional Experiments

The following results expand the main-text comparisons. A0 is standard VecSet. A1 adds the learnable decoder bank (LB), A2 adds distance bias (DB), A3 removes encoder query normals (PQ), and A4 adds normal supervision (NS). Each variant retains the preceding components.

Arrows indicate the preferred direction (↑ higher, ↓ lower). **Bold** and underlined scores mark the best and second-best per column, with ties at the displayed precision sharing a rank. Captions specify any separate comparison groups.

C.1 Reconstructing Sharp Structure

Global surface accuracy can hide errors around edges and narrow features. Table 9 therefore also reports sharp-point distance (S-CD) and sharp-point coverage (S-F). Lower S-CD means that the reconstruction lies closer to ground-truth sharp features; higher S-F means that more of those features are recovered within the threshold. The cumulative variants improve both measures, complementing the reconstruction gains in the main text.

Table 9: Sharp-structure reconstruction for A0–A4 at 4096 tokens. CD and sharp-point distance S-CD are scaled by 10^3; F1@.001 and sharp-point coverage S-F@.01 are percentages.

Model	Tokens	CD↓	F1@.001↑	S-CD↓	S-F@.01↑
A0: VecSet	4096	3.723	35.15	3.049	97.30
A1: + LB	4096	2.960	42.70	2.803	98.30
A2: + DB	4096	2.809	45.99	2.759	98.44
A3: + PQ	4096	2.723	46.71	2.480	99.11
A4: + NS	4096	**2.526**	**49.12**	**2.281**	**99.50**

C.2 Storage and Inference Cost

The shared decoder bank changes computation without enlarging the stored shape representation. For M tokens of width C_z, storage remains proportional to MC_z, while the decoder processes $M + K$ tokens. Table 10 shows that the reconstruction gains come with a modest increase in latent-refinement time relative to VecSet. Encoding, refinement, and spatial queries are shown separately to make this tradeoff visible.

Table 10: Representation storage and inference cost. KiB is stored latent size and Mem. is peak memory in GiB. Encode constructs the latent code; Decode refines it; KV prepares decoder keys and values; Query evaluates 4096 spatial locations. Times are in milliseconds and exclude mesh extraction. Blocks separate cumulative variants and full-scale models at each token budget; memory and times are ranked within each block.

Model	Tokens	KiB	Mem.↓	Encode↓	Decode↓	KV↓	Query↓
A0: VecSet	4096	512	**3.754**	**1273.9**	**53.96**	**0.466**	2.189
A1: + LB	4096	512	3.755	1276.5	56.33	0.484	2.254
A2: + DB	4096	512	3.755	1275.3	56.34	0.495	**2.146**
A3: + PQ	4096	512	3.755	1275.6	56.46	0.494	2.148
A4: + NS	4096	512	3.755	1275.5	56.49	0.497	2.175
VecSet VAE†	4096	512	**2.319**	649.7	53.99	**0.465**	3.806
LATTICE†	4096	512	2.320	**312.5**	**53.87**	**0.465**	2.187
Align-VAE	4096	512	2.320	649.8	56.32	0.493	**2.149**
VecSet VAE†	8192	1024	**3.802**	2544.2	146.04	0.898	4.958
LATTICE†	8192	1024	3.803	**670.3**	**145.82**	**0.897**	3.066
Align-VAE	8192	1024	3.803	2546.2	150.84	0.932	**2.927**

C.3 Changes Across Cumulative Variants

Table 11 measures the change between successive variants on the same objects. The decoder bank gives the clearest initial CD reduction. Distance bias and normal-free queries further improve the point estimates, while their reconstruction intervals include zero. Their spatial effects are clearer in the attention profiles below. A4 improves reconstruction while increasing angular error, reinforcing the distinction between surface fit and field orientation.

Table 11: Successive variant minus preceding variant, with 95% confidence intervals across shapes. CD is scaled by 10^3, F1 changes are percentage points, and normal-error changes are degrees. Negative CD changes and positive F1 changes indicate better reconstruction. Emphasis ranks point estimates of the changes, not statistical significance.

Comparison	Δ CD$\downarrow$	Δ F1@.001$\uparrow$	$\Delta E_{\text{angle}}\downarrow$
A0 $\rightarrow$ A1	**-0.763** [-0.958, -0.455]	**7.55** [4.83, 9.47]	**-0.91** [-1.02, -0.81]
A1 $\rightarrow$ A2	-0.150 [-0.408, 0.055]	3.29 [0.02, 5.76]	-0.31 [-0.35, -0.27]
A2 $\rightarrow$ A3	-0.087 [-0.331, 0.107]	0.72 [-2.11, 3.53]	-0.80 [-0.89, -0.71]
A3 $\rightarrow$ A4	-0.196 [-0.409, -0.007]	2.42 [0.24, 5.63]	3.66 [3.26, 4.07]

C.4 Spatial Retrieval and Query-Normal Sensitivity

The distance bias gives retrieval a spatial reference. Table 12 separates attention near the surface from attention throughout the volume. With distance bias, shape-token reads become shorter and their spatial responses more compact. Turning the bias off in A2 increases surface read distance from 0.035 to 0.607 and angular error from 11.61° to 29.81°. The anchors and latent code are unchanged, showing that the learned decoder actively uses the spatial prior during retrieval.

Table 12: Decoder attention profiles. $T_{1\%}$ is the attention share of the most-read 1% of shape tokens; H_{shape} measures how evenly shape tokens are read. G_{bank} is the shared bank's attention share. R_{shape} and R_{top} are read distances for all shape tokens and the most-read 1%, respectively; S_{sel} measures spatial response spread. Shares are percentages; distances use normalized coordinates. Distance and spread are ranked separately for surface and volume probes. Mass shares and entropy are descriptive, with no preferred direction.

Model	Probe	$T_{1\%}$	H_{shape}	G_{bank}	$R_{\text{shape}}\downarrow$	$R_{\text{top}}\downarrow$	$S_{\text{sel}}\downarrow$
A0	Surface	1.95	0.995	0.00	0.205	0.217	0.343
A0	Volume	5.00	0.985	0.00	0.874	1.197	0.921
A1	Surface	2.66	0.992	31.19	0.175	0.320	0.450
A1	Volume	3.18	0.986	38.43	0.834	0.900	0.926
A2	Surface	2.46	0.990	54.27	**0.035**	**0.028**	**0.041**
A2	Volume	2.00	0.996	97.04	**0.079**	**0.085**	**0.098**
A2 (bias off)	Surface	1.89	0.996	29.50	0.607	0.659	0.569
A2 (bias off)	Volume	1.91	0.996	30.87	1.143	1.201	1.028
A3	Surface	2.38	0.991	53.17	0.039	0.029	0.044
A3	Volume	2.32	0.994	96.40	0.093	0.105	0.108
A4	Surface	2.48	0.990	48.69	0.036	**0.028**	0.042
A4	Volume	2.23	0.994	96.24	0.081	0.092	0.103

Normal-free queries remove dependence on an unavailable cue. Zeroing query normals raises A0's CD from 4.037 to 16.414 and A2's from 3.026 to 3.747 (Table 13). In contrast, A3 retains CD 2.776 and 100% local selection under every perturbation. The encoder still observes support normals, but its queries no longer require them. This preserves orientation information while avoiding a query feature that arbitrary decoder locations cannot provide.

Table 13: Effect of changing only encoder query normals. A0 is VecSet; A2 includes the bank and distance bias; A3 additionally removes query normals. Local selection L_{enc} and local attention mass $P_{\text{enc},k}$ are percentages; CD is scaled by 10^3. Positions and support features remain fixed. Scores are ranked across models under the same query-normal condition.

Model	Query normals	Local selection ↑	Local mass ↑	CD ↓
A0	Clean	0.076	3.05	4.037
A0	Zeroed	0.052	3.04	16.414
A0	Perturbed	0.078	3.05	4.382
A0	Random flips	0.076	2.98	5.759
A2	Clean	0.043	9.77	3.026
A2	Zeroed	0.043	9.87	3.747
A2	Perturbed	0.046	9.76	3.248
A2	Random flips	0.042	9.47	4.022
A3	Clean	**100.000**	**53.82**	**2.776**
A3	Zeroed	**100.000**	**53.82**	**2.776**
A3	Perturbed	**100.000**	**53.82**	**2.776**
A3	Random flips	**100.000**	**53.82**	**2.776**

C.5 Surface Accuracy and Field Orientation

Table 14 compares the field gradient with the reference surface normal, both over the whole surface and near regions of high curvature. A3 improves orientation relative to the baseline, whereas A4 combines better reconstruction with a larger angular error. The same distinction appears near high-curvature features: the error rises from 16.84° to 18.14° between A3 and A4. A directional slope constraint therefore complements surface fitting, but does not by itself constrain every component of the field gradient.

Table 14: Reconstruction and field orientation. Angular errors are in degrees; the high-curvature subset contains points with the largest local normal variation. CD is scaled by 10^3 and sharp-point coverage S-F is a percentage. Lower angular error indicates closer agreement between predicted gradients and reference normals.

Model	Angular error ↓	High-curvature error ↓	CD ↓	S-F@.01 ↑
A0: VecSet	12.83	21.01	3.723	97.30
A1: + LB	11.92	19.09	2.960	98.30
A2: + DB	11.61	18.50	2.809	98.44
A3: + PQ	**10.82**	**16.84**	2.723	99.11
A4: + NS	14.48	18.14	**2.526**	**99.50**

D Additional Qualitative Results

Figures 8 and 9 compare seven additional selected shapes at 8192 tokens. The enlarged views illustrate local geometric detail without implying lower whole-shape reconstruction error.

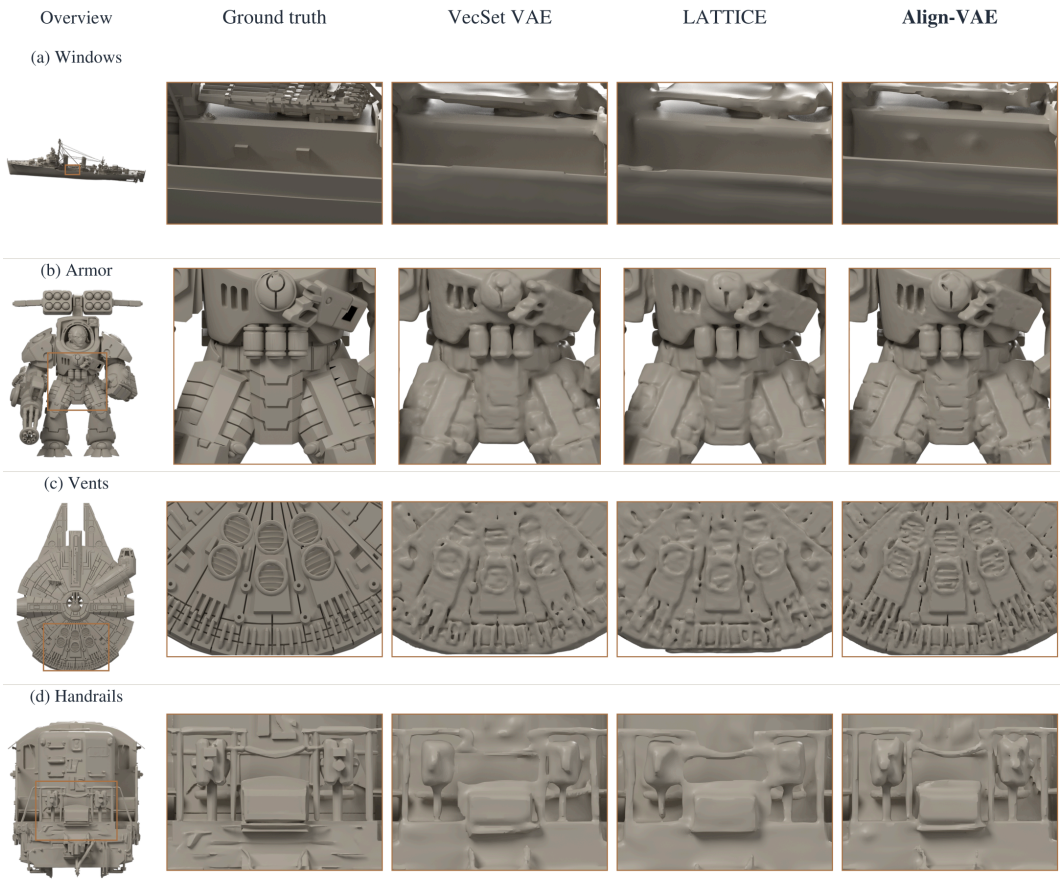


Figure 8: Additional qualitative results on selected shapes at 8192 tokens: (a) hull windows, (b) waist canisters and armor plates, (c) vents and rim struts, and (d) handrails and fixtures. Ground-truth overviews locate each detail. All methods share the camera, material, lighting, and crop within each row, with details rendered directly at higher resolution.

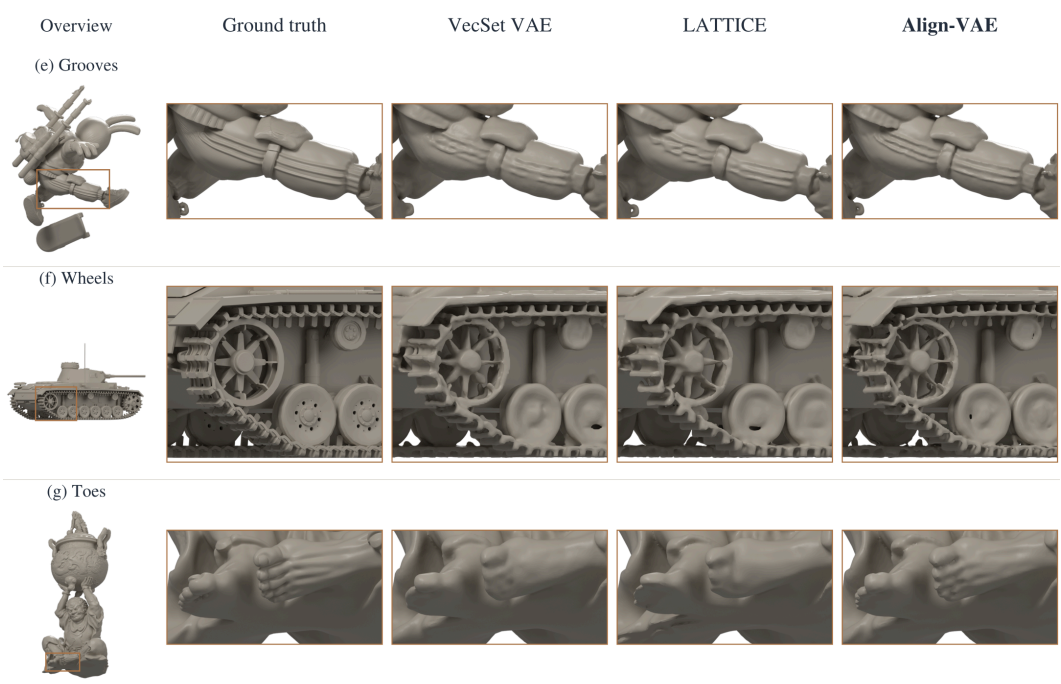


Figure 9: Additional qualitative results on selected shapes at 8192 tokens (continued): (e) trouser grooves, (f) wheel spokes and rim, and (g) individual toes. The same rendering protocol is used as in Figure 8, preserving the size of the enlarged details and the matched views across methods.