When Absence Is Not Evidence: Occlusion-Type-Aware Censored Fusion for Multi-View 3D Grounding

Anonymous authors
Paper under double-blind review

Abstract

In multi-view 3D grounding, a low language-conditioned response can indicate that a surface does not belong to the referent or that it is occluded. Conflating these cases can dilute valid evidence or accumulate spurious contradiction. We formulate fusion as *censored observation modeling*, marginalizing uncertain track–view validity. Exact visibility labels reveal a readout-dependent distinction: partial occlusion can retain semantic evidence through spatially extended response fields, whereas full occlusion carries much weaker membership signal. Our three-component likelihood preserves partial-occlusion evidence and approximates full occlusion with a shared nuisance density. A spatial gate removes instance-grouping labels from inference while geometry and semantic backbones remain frozen. We introduce VisCen3D, a diagnostic benchmark with persistent surface identity, exact visibility, and occlusion-type labels. A preregistered evaluation on 160 fresh scenes and 882 episodes shows that grouping-free fusion on analytic tracks improves AUROC on the most-occluded surface tercile by $+0.050$ under Stable Diffusion and $+0.041$ under SigLIP over development-selected non-censored comparators; scene-level 95% confidence intervals exclude zero. A post-hoc ablation on the same partition shows that the partial/full refinement improves over ordinary censoring. Top-1 target-track selection increases by $3.29/16.55$ percentage points under SD/SigLIP, with the SD directional prediction confirmatory and the SigLIP result descriptive. Transfer experiments expose the limits: averaging outperforms every tested likelihood rule on ScanNet++ with windowed responses, while incomplete surface recovery limits fusion with fully estimated geometry. These findings establish ranking and selection benefits on controlled surface support and clarify dependence on semantic readout and surface recovery.

1 Introduction

In multi-view 3D grounding, a low semantic response may indicate that a surface does not belong to the referent or that it is occluded. On an observed surface, a low response can count against membership; at a hidden projection, it may describe the occluder instead. Averaging these responses can dilute valid evidence, while accumulating uncensored likelihood ratios can produce repeated contradiction. The central question is *when is absence evidence?*

We address this question through *censored observation modeling*, marginalizing geometry-based uncertainty about whether each track–view response observes the intended surface. Exact visibility labels reveal that occluded responses need not be uninformative: when an instance remains partly visible, spatially extended response fields can carry evidence to hidden projections. Response-versus-membership AUROC is 0.76/0.80 under partial occlusion for Stable Diffusion/SigLIP, compared with 0.52/0.45 under full occlusion. This readout-dependent distinction motivates separate models for partial and full occlusion.

Our three-component likelihood separates valid, partially occluded, and fully occluded responses (Fig. 1). Class-conditional partial densities preserve informative censored responses, while a shared nuisance density approximates the weak full-occlusion signal without asserting exact label inde-

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

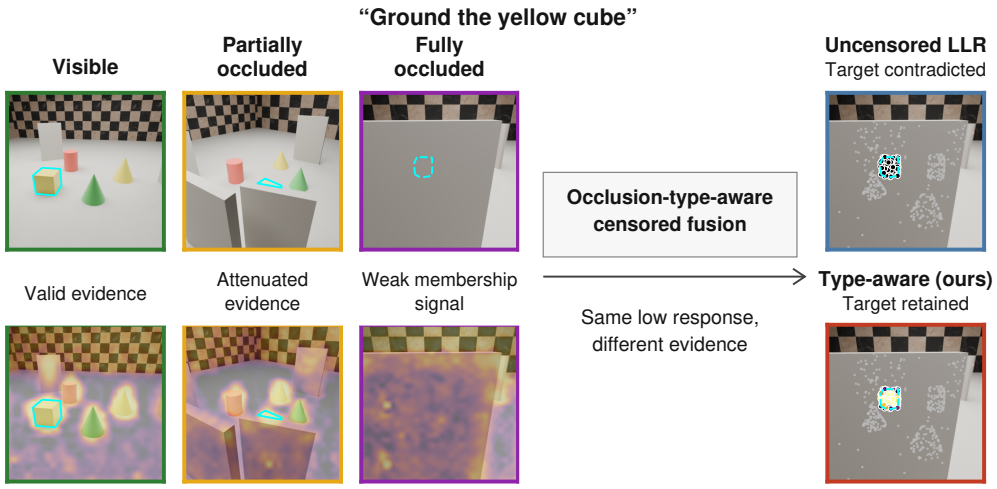

Figure 1: **Three views, three different meanings of absence.** The same target surface is visible, partially occluded, or fully occluded, with frozen Stable Diffusion responses below. Observed low responses can provide negative evidence; partial occlusion can retain an attenuated semantic halo. Weak full-occlusion signal motivates a shared nuisance density. Type-aware fusion preserves partial evidence while reducing spurious contradiction from uncensored likelihood fusion.

pendence. A fixed, scale-relative 3D neighborhood supplies the partial/full mixture gate without instance labels at inference.

We isolate fusion by fixing tracks and cached responses within each comparison, using frozen DAAM-style Stable Diffusion cross-attention (Tang et al., 2023) and SigLIP (Zhai et al., 2023). Controlled experiments use analytic tracks and estimated validity; a separate setting reconstructs support with frozen VGGT (Wang et al., 2025). Only low-capacity response models and validity calibrators are fitted on scene-disjoint development data.

We introduce *VisCen3D*, comprising 500 photorealistic synthetic scenes and 5,409 benchmark queries across eight difficulty profiles. Persistent surface identity, exact per-view validity, and analytic occlusion-type labels support controlled diagnosis, with annotations checked geometrically and through independent human verification.

On 160 fresh scenes and 882 referring episodes, a preregistered evaluation shows that grouping-free fusion on analytic tracks improves $\text{AUROC}_{\text{occ}}$ over development-selected non-censored comparators by $+0.050$ under Stable Diffusion and $+0.041$ under SigLIP; both scene-level 95% confidence intervals exclude zero. The spatial gate also meets the preregistered 0.01 ranking non-inferiority margin against an oracle-grouped gate using ground-truth instance groups and estimated validity.

A post-hoc ablation on the same partition shows that the partial/full refinement adds $+0.051/+0.061$ $\text{AUROC}_{\text{occ}}$ over the two-component censored mixture under SD/SigLIP. The complete model improves top-1 target-track selection over the frozen comparators by $3.29/16.55$ percentage points, with both paired 95% confidence intervals excluding zero. The SD directional prediction is confirmatory; the SigLIP selection result remains descriptive.

Transfer experiments expose the limits of these gains. On ScanNet++ (Yeshwanth et al., 2023), averaging outperforms every tested likelihood rule with windowed responses. With fully estimated geometry, reconstruction reaches only 0.39 of sampled target tracks overall, so strong ranking on recovered support does not establish performance over the full target surface. The benefits therefore depend on semantic readout and surface recovery.

Contributions. (1) We model multi-view semantic evidence through censored observations, accounting for uncertain surface validity with frozen backbones. (2) We distinguish partial from full occlusion using a three-component likelihood and a grouping-free spatial gate. (3) We introduce VisCen3D, with persistent surface identity, exact validity and occlusion-type labels, controlled dif-

ficulty, and verified queries. (4) On fresh synthetic data with analytic support, we establish preregistered ranking gains, demonstrate the partial/full refinement's benefit in a post-hoc grouping-free ablation, and improve target-track selection. Transfer tests identify limits from semantic readout and surface recovery.

2 Related Work

Multi-view geometry and 3D grounding. Closest to our sparse-view setting is MVGGT (Wu et al., 2026), which segments referred objects from RGB views using frozen reconstruction and trainable multimodal branches. Its Per-view No-target Suppression Optimization (PVSO) balances supervision from target-visible and no-target views. This view-level training objective does not specify how a projected surface response contributes to a likelihood at inference. SeeGround (Li et al., 2025) renders a query-relevant viewpoint for zero-shot 3D localization, while ReasonGrounder (Liu et al., 2025) uses hierarchical Gaussian features and vision–language-model guidance to ground objects, including occluded ones. We study the evidence carried by a projected response when observation of its surface is uncertain.

2D-to-3D semantic lifting and open-vocabulary perception. MaskClustering (Yan et al., 2024) groups 2D masks by multi-view consensus, Open3DIS (Nguyen et al., 2024) uses them to guide 3D instance proposals, and Any3DIS (Nguyen et al., 2025) tracks masks and refines the resulting proposals. ConceptGraphs (Gu et al., 2024) associates foundation-model outputs into an open-vocabulary 3D scene graph. Gaussian-based approaches include LangSplat's language field (Qin et al., 2024), Lifting by Gaussians' instance segmentation (Chacko et al., 2025), and Dr. Splat's direct language-feature registration (Kim et al., 2025). GenSplat (Liu et al., 2026) studies cross-scene language generalization and frame selection, while ReLaGS (Xie et al., 2026) aligns multi-view language features and reasons over relations. These methods build semantic 3D representations or proposals; we hold support and responses fixed to isolate fusion under uncertain visibility.

Dense semantic readouts and grounding benchmarks. DAAM (Tang et al., 2023) extracts attribution maps from Stable Diffusion cross-attention, while ConceptAttention (Helbling et al., 2025) derives concept maps from diffusion-transformer attention. We also localize SigLIP's image–text representations (Zhai et al., 2023) through frozen sliding-window readout. The measured partial-occlusion signal therefore depends on the readout as well as geometry.

ScanRefer (Chen et al., 2020) and ReferIt3D's Nr3D/Sr3D datasets (Achlioptas et al., 2020) provide object-level grounding annotations on ScanNet (Dai et al., 2017). MVGGT's MVRefer (Wu et al., 2026) evaluates sparse-view referring segmentation, while ReasonGrounder's ReasoningGD (Liu et al., 2025) includes modal and amodal masks for occluded objects. These benchmarks do not jointly provide persistent surface identity, per-surface validity, occluder identity, and explicit partial-versus-full external occlusion. VisCen3D supplies this combination to diagnose each projected response's contribution.

Visibility-aware fusion and censored inference. VisFusion (Gao et al., 2023) estimates voxel visibility to guide image-feature aggregation for reconstruction. We use estimated observation validity to determine the semantic evidence assigned to an extracted response.

As in missing-data inference (Rubin, 1976; Little & Rubin, 2019), the likelihood accounts for the observation mechanism. Here, however, a hidden projection still yields a response that may reflect an occluder or a spatially extended semantic field. Our measurements motivate class-conditional partial-occlusion densities and a shared full-occlusion density. The latter approximates weak membership signal without assuming exact independence, preserving partial evidence while reducing spurious contradiction.

3 Method

3.1 Surface tracks and validity

Let m index a referring mention, t a physical surface track with canonical position $\mathbf{x}_t \in \mathbb{R}^3$, and i a view. The set $\mathcal{I}_t$ contains views in which t projects in-frustum, at image location u_{ti}. We distin-

guish *track support*, which specifies the represented surfaces and their projections, from *observation validity*, which asks whether the image at u_{ti} observes surface t rather than an occluder.

Controlled experiments use *analytic tracks*: ground-truth geometry provides track identities, canonical positions, and projections, while observation validity remains estimated. This isolates fusion from surface-recovery and correspondence errors. In the *fully estimated* setting, frozen VGGT (Wang et al., 2025) supplies tracks, positions, projections, and geometric cues. No ground-truth camera, depth, surface geometry, or instance grouping enters its test-time inference path; ground truth is used only for evaluation.

For $i \in \mathcal{I}_t$, let $A_{ti} \in \{0, 1\}$ indicate whether the projected sample observes surface t. A low-capacity logistic predictor followed by histogram calibration maps relative depth margin, depth confidence, and point-map residual to

$$\eta_{ti} = \widehat{P}(A_{ti} = 1 \mid \text{geometric cues}). \tag{1}$$

The predictor and calibrator are fitted on scene-disjoint development data and then frozen. Because η_{ti} depends only on geometry, it is shared across mentions and referent hypotheses; inference never uses ground-truth A_{ti}. Off-frustum pairs are masked from fusion rather than treated as observed low responses. Implementation and calibration details appear in App. B.1.

3.2 Censored observation model

Let $Z_{tm} \in \{0, 1\}$ indicate referent membership and $R_{tim} \in [0, 1]$ the normalized semantic response at u_{ti}. Development data provide visible-response densities $f_z(r) = p(R_{tim} = r \mid Z_{tm} = z, A_{ti} = 1)$ and a censored nuisance density $c(r)$ shared across membership hypotheses. Marginalizing validity gives

$$\begin{aligned} q_z^{\text{cens}}(r \mid \eta) &= \eta f_z(r) + (1 - \eta)c(r), \\ d_{tim} = \beta_{tim} \log \frac{q_1^{\text{cens}}(R_{tim} \mid \eta_{ti})}{q_0^{\text{cens}}(R_{tim} \mid \eta_{ti})}, &\qquad \pi_{tm} = \sigma\left(\lambda + \sum_{i \in \mathcal{I}_t} d_{tim}\right). \end{aligned} \tag{2}$$

Here λ is a development-frozen prior log odds and π_{tm} the track score. The validity weight η_{ti} estimates whether the surface was observed; the reliability weight β_{tim} tempers uncertain semantic readouts. For SD, $\beta_{tim} = \text{clip}((1 + 0.5 s_{tim})^{-1}, 0.05, 1)$, where s_{tim} measures disagreement across six standardized readouts. SigLIP has no analogous ensemble and uses $\beta_{tim} = 1$ in-frustum.

At $\eta = 0$, both hypotheses reduce to c and contribute zero evidence; at $\eta = 1$, the message is the visible log-likelihood ratio scaled by β_{tim}. A low response therefore provides negative evidence only when the fitted likelihood ratio is below one. Estimated validity, pooled densities, and correlated views preclude an exact posterior interpretation. We treat the sum as a reliability-weighted composite log-likelihood ratio and evaluate calibration separately (App. C.3).

3.3 Occlusion-type-aware censoring

A censored surface is *partially occluded* when another surface of its physical instance remains visible in the view, and *fully occluded* otherwise. Spatially extended response fields can preserve semantic evidence at hidden projections under partial occlusion. Response-versus-membership AUROC is 0.76/0.80 under partial occlusion, compared with 0.52/0.45 under full occlusion for SD/SigLIP. The latter supports a weak-signal approximation, not exact independence.

We retain class-conditional partial densities $g_z(r) = p(R_{tim} = r \mid Z_{tm} = z, A_{ti} = 0, \text{partial})$ and model full occlusion with a shared density $h(r)$. A mixture gate $\rho_{ti} \in [0, 1]$ divides the censored branch:

$$q_z^{\text{type}}(r \mid \eta, \rho) = \eta f_z(r) + (1 - \eta)\big[\rho\, g_z(r) + (1 - \rho)h(r)\big]. \tag{3}$$

Fusion substitutes q^{type} into Eq. (2). At $\eta = 1$, the likelihood ratio is f_1/f_0; at $\eta = 0, \rho = 1$, it is g_1/g_0; and at $\eta = 0, \rho = 0$, class evidence is exactly zero by construction. Estimated gates need not reach this last endpoint on fully occluded pairs, so empirical contradiction can remain nonzero. Because c pools all censored observations whereas h models full occlusion, setting $\rho = 0$ recovers the separately fitted two-component baseline only if $h = c$.

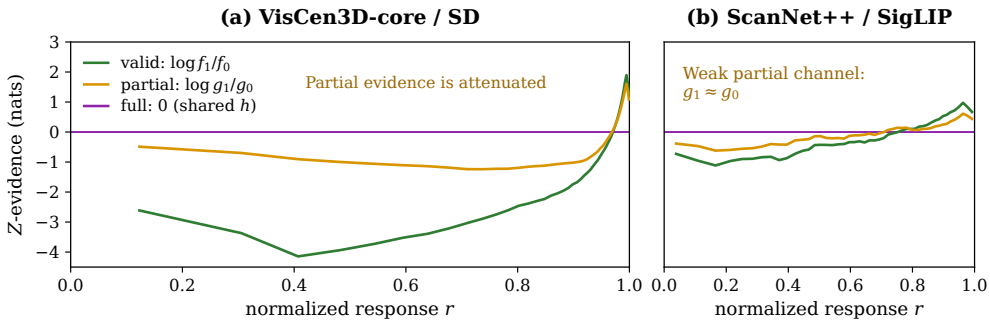

Figure 2: **Measured partial-occlusion evidence.** Development-fitted log-likelihood ratios for visible and partially occluded responses on (a) VisCen3D-core with SD and (b) ScanNet++ with SigLIP. Both panels show the full curve range on a shared vertical scale. Partial responses retain attenuated visible evidence on core but provide weak evidence on ScanNet++. The purple zero line follows from the shared-h full-occlusion model, not an empirical equality. A five-bin moving average is used for visualization only.

Grouping-free gate. Let $D = \|\max_t \mathbf{x}_t - \min_t \mathbf{x}_t\|_2 + 10^{-6}$ be the track set's bounding-box diagonal with a numerical safeguard, where extrema are coordinatewise. With fixed $\gamma = 0.06$, define

$$\mathcal{N}_\gamma(t) = \{t' : \|\mathbf{x}_{t'} - \mathbf{x}_t\|_2 \le \gamma D\}, \qquad \rho^{\mathrm{sp}}_{ti} = \max_{t' \in \mathcal{N}_\gamma(t)} \eta_{t'i}. \tag{4}$$

The neighborhood includes the scored track, so it is nonempty and $\rho^{\mathrm{sp}}_{ti} \ge \eta_{ti}$. This gate summarizes estimated visibility in a local neighborhood; it is a geometric mixture weight, not a calibrated partial-occlusion probability. Analytic experiments use ground-truth track positions, whereas the fully estimated setting uses VGGT positions.

For diagnosis, an *oracle-grouped gate* replaces $\mathcal{N}_\gamma(t)$ with the ground-truth instance group $\mathcal{G}(t)$, also including t. Both gates use estimated validity. Non-inferiority to this reference assesses the effect of substituting the spatial gate on fusion performance, not occlusion-type classification accuracy. The primary spatial gate removes instance-grouping labels from inference; analytic track support and labeled development data remain part of the controlled setting.

3.4 Evaluation targets

An object-referring episode is a scene–mention pair with track labels Z_{tm}. $\mathrm{AUROC}_{\mathrm{occ}}$ is computed within each episode's tie-preserving most-occluded tercile, defined by the number of truly valid views, and averaged across episodes; single-class terciles are excluded. R@1-track selects the highest-scoring track among all episode tracks and records whether it belongs to the target instance, resolving score ties by the first track index. This measures target-track selection on the available support, rather than localization or segmentation. Ground-truth visibility and membership define these metrics but are not used by the inference gates.

Estimated-geometry results distinguish ranking on matched recovered support from a reference-track factorial that retains unreached surfaces. Exact tercile construction, realized populations, matching, and missing-score rules appear in Apps. B.5 and E.

4 The VisCen3D Benchmark

Existing 3D grounding benchmarks supervise objects or view-level targets, whereas our observation model requires *surface-level* supervision: whether a response sampled at a projected location observes the intended physical surface. VisCen3D provides persistent surface identity, per-view validity, and occlusion type, with controlled mixtures of visible and occluded observations. This supports fusion comparisons with fixed geometric support and semantic responses.

Synthetic core. The core contains 500 photorealistic scenes rendered with Blender Cycles from procedural geometry and CC0 materials, without external object assets or images of people. Eight difficulty profiles cover heavy occlusion, low-overlap camera clusters, touching and identical objects, small or thin objects, and stacked layouts requiring gravity-frame relations. The 5,409 benchmark query episodes include attribute, relational, and null queries, with explicit reference frames, unambiguity margins, paraphrase groups, and visibility-stratum labels. These query counts differ from the object-referring scene–mention pairs used to evaluate fusion; evaluation populations are specified in Sec. 5 and App. Table 3.

Surface identity and visibility. Ground truth is derived from geometry independently of RGB appearance. Each track stores persistent barycentric coordinates on its source triangle. A projection is valid only when a z-buffer identity test and an independent nearest-hit ray cast agree, within tolerance, on the same front-facing surface. Their agreement is 0.999; disagreements are excluded from visibility labels. The occlusion analysis considers in-frustum, front-facing observations under *external occlusion*, recording the occluding instance. Partial versus full occlusion depends on whether any other surface of the track's own instance remains visible in the view, irrespective of the referring mention (Sec. 3.3).

Appearance and transfer. The synthetic core has two renderings with identical geometry, ground truth, queries, and splits. We use `core-agx-reference` for all reported synthetic results and human verification. The alternative `core-standard` improves color fidelity; results from the two renderings are not pooled.

The Replica transfer partition (Straub et al., 2019) spans all 18 environments, with 58 region-scenes and 456 episodes split by environment. It combines baked scan appearance with semantic-mesh geometry and the same visibility construction; the two geometric tests agree at 0.994. These construction totals are distinct from the realized test populations after quality control and metric-specific exclusions (Sec. 5.3). ScanNet++ adds real DSLR photographs, providing a further test of appearance shift.

Human verification and release. Two raters independently re-annotated all 3,506 benchmark queries in the 324-scene core test split, achieving target-identity agreement of 0.980 and Cohen's $\kappa = 0.975$. A separate adjudicator reviewed disagreements; unresolved cases concentrate in touching, same-class configurations. Construction, color fidelity, quality assurance, annotation, and release details appear in App. A. Synthetic renders, surface-level ground truth, queries, splits, and evaluation metadata will be separately released under CC BY 4.0. For Replica and ScanNet++, we provide reconstruction code and metadata for independently obtained datasets rather than redistributing their imagery or geometry-derived arrays.

5 Experiments

Protocol. We hold tracks and cached semantic responses fixed within each comparison. Densities, validity calibration, and thresholds are fitted on scene-disjoint development data. Confirmation 2 provides the primary grouping-free test, with preregistered analysis rules and development-frozen comparators; its additional two-component ablation is post-hoc. Repeatedly inspected core results are descriptive. Paired 95% CIs use scene-level bootstrap resampling. Confirmation 1 has a documented command clarification preceding its single sealed evaluation (App. B.2). Populations, metrics, and exclusion rules appear in App. B.

5.1 Grouping-free sealed confirmation

Primary ranking result. Table 1 compares the spatial type-aware model with each backbone's development-selected non-censored comparator on 160 fresh scenes and 882 episodes per backbone, with no exclusions. Both paired CIs exclude zero, satisfying the preregistered conjunctive criterion. Inference uses analytic surface tracks, estimated validity, and the fixed spatial gate, without instance-grouping labels.

Against the oracle-grouped gate, which uses ground-truth instance groups and estimated validity, spatial-minus-grouped differences are $-0.0002/-0.0057$ under SD/SigLIP, with one-sided 95%

Backbone	Frozen comparator	$\text{AUROC}_{\text{occ}} \uparrow$ Comparator	$\text{AUROC}_{\text{occ}} \uparrow$ Spatial gate	$\Delta\text{AUROC}_{\text{occ}} \uparrow$	95% CI
SD 1.5	η-weighted LLR	0.818 ± 0.15	$\mathbf{0.868} \pm 0.14$	$+\mathbf{0.0503}$	$[+0.0453, +0.0552]$
SigLIP	Mean, all views	0.885 ± 0.17	$\mathbf{0.927} \pm 0.12$	$+\mathbf{0.0414}$	$[+0.0318, +0.0513]$

Table 1: **Primary sealed evaluation of grouping-free fusion.** Results on 160 scenes and 882 episodes per backbone, with analytic tracks, estimated validity, and no exclusions. Absolute levels are descriptive post-hoc summaries of sealed scores (episode mean $\pm$ standard deviation); registered inference uses paired gains and scene-bootstrap 95% CIs. Gains use unrounded scores. The spatial gate requires no instance labels at inference.

Backbone	Frozen comparator	R@1-track (%)↑ Comparator	R@1-track (%)↑ Spatial gate	Gain (pp)	95% CI (pp)
SD 1.5	η-weighted LLR	63.95	$\mathbf{67.23}$	$+\mathbf{3.29}$	$[+0.46, +6.00]$
SigLIP	Mean, all views	57.48	$\mathbf{74.04}$	$+\mathbf{16.55}$	$[+13.03, +20.07]$

Table 2: **Target-track selection on the confirmation partition.** R@1-track measures selection of a target track among all episode tracks, using the same 160 scenes and 882 episodes per backbone without exclusions. Gains and paired scene-bootstrap 95% CIs use unrounded outcomes; pp denotes percentage points. Intervals were computed in the original sealed run. The SD directional prediction is registered; SigLIP is descriptive.

lower bounds of $-0.0024/-0.0074$. Both exceed the preregistered -0.01 margin, establishing ranking non-inferiority at that margin. A separate sealed confirmation of the oracle-grouped model appears in App. B.2.

Contribution of the partial/full refinement. A post-hoc ablation on the same episodes compares the spatial type-aware model with the two-component censored mixture, holding cached responses, validity estimates, and development-fitted densities fixed. The refinement adds $+0.0510$ $\text{AUROC}_{\text{occ}}$ (95% CI: $[+0.0452, +0.0569]$) under SD and $+0.0608$ ($[+0.0556, +0.0661]$) under SigLIP. The two-component mixture reaches approximately $0.817/0.866$, compared with $0.818/0.885$ for the frozen comparators. The refinement therefore supplies the observed advantage over these rules in the tested configurations.

Target-track selection. Table 2 evaluates whether the highest-scoring track among all episode tracks belongs to the referred object. R@1-track improves by 3.29 percentage points under SD and 16.55 under SigLIP, corresponding to 29 and 146 additional correct selections across 882 episodes. Both paired CIs exclude zero. The SD result satisfies the registered directional prediction; the SigLIP result retains its prespecified descriptive status. These results measure selection on analytic surface support rather than complete object localization or segmentation.

5.2 What drives the gain?

Occlusion diagnostics. Supporting diagnostics use 320 core scenes (40 per difficulty profile) and 1,762 episodes per backbone, with analytic tracks, estimated validity, and the diagnostic oracle-grouped gate. All episodes have defined $\text{AUROC}_{\text{occ}}$. The type-aware mixture reaches 0.873/0.934 under SD/SigLIP and outperforms the two-component mixture. All eight profile-wise gains over the pooled non-censored comparators are positive under both backbones (16 of 16 survive Holm–Bonferroni correction). These core comparisons remain descriptive (App. B.4).

Uncensored LLR contributes $1.652/3.539$ nats of mean negative evidence on fully occluded pairs under SD/SigLIP. The two-component mixture reduces this to $0.001/0.002$ nats; the type-aware mixture yields $0.005/0.011$ nats while improving ranking by retaining partial-occlusion evidence. The refinement therefore trades a small increase in residual negative evidence for better discrimination. This one-sided diagnostic pools target and non-target pairs and measures negative-message suppression; ranking and selection metrics assess its usefulness for grounding.

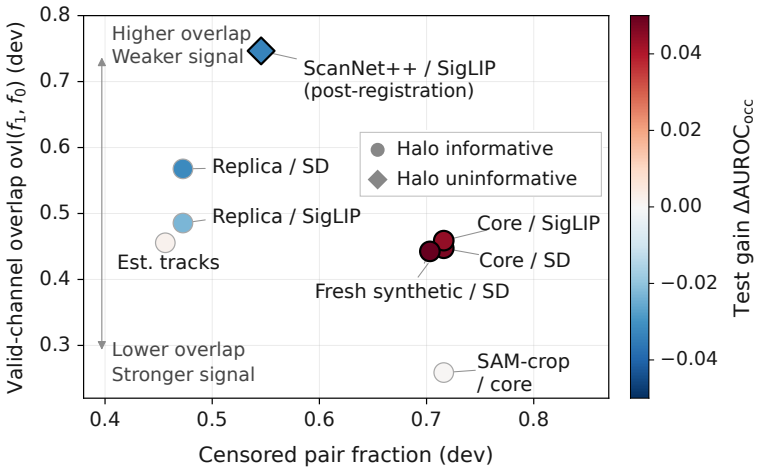

Figure 3: **Measured regimes of censoring gain.** Development-measured censored fraction and valid-channel overlap locate 8 regimes; lower overlap indicates a more informative valid channel. Color encodes test $\Delta\mathrm{AUROC}_{\mathrm{occ}}$ over the strongest non-censored rule. Bold borders mark 95% CIs excluding zero; diamonds denote weak partial channels ($\mathrm{ovl}(g_1, g_0) > 0.7$), and circles denote informative partial channels. The registered ScanNet++ parity prediction was refuted (-0.036 $[-0.051, -0.021]$): these coordinates describe regimes but do not suffice to select a fusion rule.

Reliability, validity, and learned fusion. Under SD, setting $\beta = 1$ adds 0.015 $\mathrm{AUROC}_{\mathrm{occ}}$ but doubles full-occlusion negative evidence and worsens calibration. Replacing the multi-cue validity model with a calibrated depth-margin cue changes $\mathrm{AUROC}_{\mathrm{occ}}$ by only about -0.001 under either backbone. Dependence rescaling improves SigLIP ECE but worsens SD ECE, consistent with interpreting the scores as composite-likelihood ranking functions. An additive MLP is competitive but seed-sensitive at full supervision. With one development scene, the mixture leads by mean gains of $+0.032/+0.063$ $\mathrm{AUROC}_{\mathrm{occ}}$ under SigLIP/SD across three fittable subsets out of five; two leave a density stratum empty. The MLP becomes competitive after roughly five to ten scenes, supporting a low-data advantage conditional on successful fitting (App. D).

5.3 Geometry and Appearance Shifts

Fully estimated geometry. With VGGT-estimated tracks, projections, positions, and validity cues, the type-aware model reaches 0.931 $\mathrm{AUROC}_{\mathrm{occ}}$ under SigLIP on matched estimated tracks. This result is conditional on recovered support: the paired analytic-track comparison uses different surface populations ($404/672$ episodes across $109/120$ scenes) and does not establish common-support equivalence. Target-track recall is 0.39 over all target tracks. A separate coverage-aware factorial retains a fixed ground-truth occluded tercile, assigning unreached positive and negative tracks a shared bottom rank. Its minimum AUROC across settings is 0.54, and fusion rules differ by only about 0.003. Grid seeding under-samples heavily occluded regions by 2.0–4.7×, identifying surface recovery as the dominant limitation in this regime (App. E).

Replica. Replica introduces scanned appearance and reconstruction shift while retaining exact semantic-mesh geometry for evaluation. After a pose-supervised alignment gate, 116 episodes across 38 scenes have defined $\mathrm{AUROC}_{\mathrm{occ}}$. Contrasts with the strongest non-censored rule are -0.024 $[-0.056, +0.006]$ under SigLIP and -0.034 $[-0.070, +0.002]$ under SD; both intervals include zero. Mean negative evidence on fully occluded pairs falls from 1.42 nats under uncensored LLR to 0.11 nats under the type-aware model. Negative-message suppression transfers, while a ranking gain remains unresolved.

Registered ScanNet++ test. Figure 3 shows gains concentrated where censoring is frequent and the valid semantic channel is informative. ScanNet++ refutes the registered parity prediction: with windowed responses, the paired difference from the strongest non-censored rule is -0.036 (95%

CI: $[-0.051, -0.021]$). Mean-all beats every tested likelihood rule (0.762 versus 0.726 for ours). Within the likelihood family, type-aware fusion improves over uncensored LLR by +0.026, including a $+0.009$ $[+0.005, +0.014]$ gain from the partial/full refinement.

On the same 900 episodes, mask-crop responses reverse the relative ordering: type-aware fusion reaches 0.696 versus 0.544 for mean-all, exceeding the strongest non-censored rule by $+0.021$ $[+0.016, +0.027]$. Both absolute AUROCs are lower than their window-readout counterparts. This reversal implicates a readout–response-model interaction, without isolating density-estimation error from other readout effects. Validity AUROC is also only 0.766, below the predeclared 0.90 gate. The result supports the partial/full distinction within the tested likelihood family while exposing limits of the readout, fitted model, and estimated validity (App. F.3).

5.4 External training-free pipelines

We compare frozen-SigLIP reimplementations of OpenScene-style feature averaging (Peng et al., 2023) and MaskClustering/OVIR-3D-style mask lifting (Yan et al., 2024; Lu et al., 2023). On common analytic support (120 core scenes, 672 episodes), our window-readout mixture exceeds feature averaging by $+0.061$ $[+0.052, +0.069]$, while mask lifting exceeds the window-readout mixture by $+0.019$ $[+0.010, +0.028]$. Applying type-aware fusion to mask-crop scores exceeds mask retrieval by $+0.012$ $[+0.007, +0.018]$. On Replica, the window-readout contrasts include zero and the crop-readout mixture trails mask lifting (-0.114 $[-0.163, -0.064]$). Object-level readouts and censored fusion are therefore complementary on core, but their combination does not transfer uniformly (App. G).

6 Limitations

Incomplete surface recovery limits deployment. The primary sealed confirmation uses analytic tracks; removing instance grouping at inference does not recover missing surfaces. VGGT recovers only 0.39 of sampled target tracks overall—neither occluded-only nor area-weighted recall—and disproportionately misses heavily occluded regions. Estimated-track AUROC is therefore conditional on recovered support. Surface discovery remains outside our observation model, and the estimated pipeline inherits VGGT's static-scene assumptions.

Gains also depend on the semantic readout and fitted response model. The ScanNet++ readout reversal demonstrates this interaction without isolating density misspecification. When $g_1 \approx g_0$, the partial branch adds little beyond ordinary censoring; inaccurate validity further limits fusion. Correlated views make the score a composite log-likelihood ratio, so improved ranking need not imply calibration (App. C.3).

Validity and response models require labeled development data: *grouping-free* describes inference, not supervision-free estimation. Touching identical instances and null-target abstention remain unresolved. Core diagnostics are descriptive, while the principal ranking gains are confirmed on fresh partitions from the same synthetic process. Real-data evidence is limited to Replica and ScanNet++, including the failed registered ScanNet++ parity prediction; population and exclusion details appear in App. B.5.

7 Conclusion

We formulate multi-view fusion as censored observation modeling, distinguishing informative partial occlusion from weak full-occlusion signal. On analytic support, preregistered evaluation confirms grouping-free ranking gains; a post-hoc ablation supports the partial/full refinement. Target-track selection also improves, with SD confirmatory and SigLIP descriptive. VisCen3D enables these tests through persistent surface identity and exact validity and occlusion-type labels. Transfer results identify semantic readout and surface recovery as key limitations. Future work will explore geometry-aware surface discovery and readout-adaptive response models, extending evaluation from recovered tracks to object-level grounding with fully estimated geometry. A further direction is active view selection: observation uncertainty could guide new viewpoints toward poorly observed surfaces, testing whether better evidence acquisition improves grounding beyond fusion of a fixed view set.

AI USE STATEMENT

In preparing this work, the authors used OpenAI and Anthropic language models to assist with manuscript editing and reference checking, figure and table preparation, experiment scripts, and supplementary documentation. The tools also provided feedback on experimental design, presentation, and interpretation of results. The authors reviewed and revised the assisted work, checked references and reported results against their source records, and take full responsibility for the methods, data, claims, and final manuscript.

ETHICS STATEMENT

This work studies multi-view 3D grounding in photorealistic synthetic scenes. The benchmark contains no images of people or personally identifiable information. Human verification consists of judgments about rendered scenes and does not require sensitive personal data. Asset provenance and the applicable release terms for benchmark components are documented separately. The method may inform embodied and robotic perception, where grounding errors could affect downstream decisions; we do not evaluate it for safety-critical deployment.

REPRODUCIBILITY STATEMENT

Section 3 specifies the observation model and inference gates; Sec. 4 describes benchmark construction. Appendices B.1, B.5, D, and E document implementation, statistical analysis, learned-baseline training, and estimated-geometry evaluation. The anonymous supplementary ZIP provides frozen confirmation protocols, partition manifests, benchmark metadata, real-data builder scripts, and Confirmation 2 result records, including the post-hoc ablation and selection verification. Saved scene-level aggregates and bootstrap settings permit recomputation of the reported paired means and confidence intervals. Full prediction reproduction additionally requires inference code, rendered views, cached responses, and geometric inputs outside this package. Synthetic data will be released separately; Replica and ScanNet++ inputs require independently obtained access. The supplement's README and provenance documentation specify the included artifacts, dependencies, and verification scope.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *ECCV*, pp. 422–440, 2020. doi: 10.1007/978-3-030-58452-8_25.

Rohan Chacko, Nicolai Häni, Eldar Khaliullin, Lin Sun, and Douglas Lee. Lifting by Gaussians: A simple, fast and flexible method for 3D instance segmentation. In *WACV*, pp. 3497–3507, 2025. doi: 10.1109/WACV61041.2025.00345.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *ECCV*, pp. 202–221, 2020. doi: 10.1007/978-3-030-58565-5_13.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017.

Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. BlenderProc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. doi: 10.21105/joss.04901.

Huiyu Gao, Wei Mao, and Miaomiao Liu. VisFusion: Visibility-aware online 3D scene reconstruction from videos. In *CVPR*, pp. 17317–17326, 2023.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan,

Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *ICRA*, pp. 5021–5028, 2024. doi: 10.1109/ICRA57147.2024.10610243.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion transformers learn highly interpretable features. In *ICML*, pp. 22946–22963, 2025.

Jun-Seong Kim, GeonU Kim, Yu-Ji Kim, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. Splat: Directly referring 3D Gaussian splatting via direct language embedding registration. In *CVPR*, pp. 14137–14146, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. SeeGround: See and ground for zero-shot open-vocabulary 3D visual grounding. In *CVPR*, pp. 3707–3717, 2025.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 3rd edition, 2019. doi: 10.1002/9781119482260.

Fang Liu, Yuhao Liu, Ke Xu, Gerhard Petrus Hancke, and Rynson W. H. Lau. GenSplat: Bridging the generalization gap in 3DGS language comprehension. In *CVPR*, pp. 5221–5231, 2026.

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. ReasonGrounder: LVLM-guided hierarchical feature splatting for open-vocabulary 3D visual grounding and reasoning. In *CVPR*, pp. 3718–3727, 2025.

Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. OVIR-3D: Open-vocabulary 3D instance retrieval without training on 3D data. In *CoRL*, pp. 1610–1620, 2023.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. In *CVPR*, pp. 4018–4028, 2024.

Phuc Nguyen, Minh Luu, Anh Tran, Cuong Pham, and Khoi Nguyen. Any3DIS: Class-agnostic 3D instance segmentation by 2D mask tracking. In *CVPR*, pp. 3636–3645, 2025.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, pp. 815–824, 2023.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language Gaussian splatting. In *CVPR*, pp. 20051–20060, 2024.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting Stable Diffusion using cross attention. In *ACL*, pp. 5644–5659, 2023. doi: 10.18653/v1/2023.acl-long.310.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025.

Changli Wu, Haodong Wang, Jiayi Ji, Yutian Yao, Chunsai Du, Jihua Kang, Yanwei Fu, and Liujuan Cao. MVGGT: Multimodal visual geometry grounded transformer for multiview 3D referring expression segmentation. In *CVPR*, pp. 16800–16810, 2026.

Yaxu Xie, Abdalla Arafa, Alireza Javanmardi, Christen Millerdurai, Jia Cheng Hu, Shaoxiang Wang, Alain Pagani, and Didier Stricker. ReLaGS: Relational language Gaussian splatting. In *CVPR*, pp. 23826–23836, 2026.

Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. MaskClustering: View consensus based mask graph clustering for open-vocabulary 3D instance segmentation. In *CVPR*, pp. 28274–28284, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *ICCV*, pp. 12–22, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023.

A BENCHMARK CONSTRUCTION AND EVALUATION POPULATIONS

Scene generation. VisCen3D samples object counts, geometry, materials, occluders, and camera trajectories from difficulty-specific procedural distributions. Blender Cycles renders the images following the general BlenderProc2 pipeline (Denninger et al., 2023). Materials are CC0; scenes contain no external object assets or images of people.

Surface identity and visibility. Each track stores barycentric coordinates on its source triangle, preserving surface identity across views. Visibility is independent of RGB appearance: z-buffer identity and a nearest-hit ray cast must agree, within tolerance, on the same front-facing surface. Agreement is 0.999; disagreements receive no visibility label. Each censored observation records its occluding instance. An occlusion is partial if the track's physical instance has any visible pixels in that view, and full otherwise. Automated checks cover camera validity, duplicate surface identifiers, z-buffer/ray consistency, query executability, visibility mixtures, and stratum balance.

Color fidelity. The reference core uses Blender's AgX view transform, with an sRGB palette originally supplied to Base Color as if linear. Across 500 scenes, rendered HSV saturation is 36–55% of the specified value, with hue shifts up to $12°$. Orange–yellow separation decreases from $23.5°$ to $15.2°$; red–orange separation remains $25.8°$. Of 70 primary-rater disagreements, 45 concern competing objects and 31 involve same-color objects. The remaining 14 cross-color disagreements represent approximately 0.4% of the test query episodes.

Both renderings share analytic ground truth, episodes, and splits. All reported results and human verification use `core-agx-reference`. The corrected `core-standard` uses the Standard view transform and a linearized palette, retaining 78–84% of specified saturation, with hue shifts below $6°$ and orange–yellow separation of $23.2°$. Results are not pooled across renderings.

Human verification. Two raters independently re-annotated all 3,506 benchmark test queries using the rendered images and referring expression. Candidate order was independently shuffled; neither rater saw the stored target or the other's response. Target-identity agreement is 0.980 (Cohen's $\kappa = 0.975$), and relational-query agreement is 0.973, exceeding the preregistered $0.95/0.90$ gates. Agreement with the analytic target is 0.985 and 0.991 for the two raters.

Agreement is lowest in the same-class stratum (0.933) and touching configurations (0.938), and is 0.978 in the heavy-occlusion stratum. A separate adjudicator reviewed the 70 disagreements with freshly shuffled candidates and no access to either original response or the analytic target. The adjudicator resolved 54 cases and judged 16 ambiguous, yielding a human-resolved fraction of 0.995. The unresolved cases concentrate in touching, same-class scenes. Among resolved disagreements, agreement with the analytic target is 0.981. Neither primary rater flagged unnatural English; the adjudicator flagged 2 expressions among the disputed cases.

Evaluation units and subsets. Benchmark episodes include attribute, relational, and null queries. Controlled fusion uses unique scene–mention pairs for unambiguous non-occluder objects, excluding ambiguous twins. Table 3 separates these populations. The full test split also has four baseline scenes with 34 benchmark queries excluded from controlled fusion.

Real-data partitions and release scope. Each benchmark query records its scene, referring expression, unambiguity margin, paraphrase group, directional reference frame, and visibility stratum. Replica uses scanned semantic meshes for geometry, depth, and segmentation, with appearance rendered from baked textures. Environment-disjoint splits prevent overlap between development and test environments. ScanNet++ uses real DSLR photographs at registered poses.

We release builder code, scene selections, splits, episode text, and integrity metadata for reconstruction from independently obtained access to Replica and ScanNet++. We do not redistribute their imagery, depth, segmentation, geometry-derived buffers, or cached arrays. Synthetic VisCen3D renders, analytic ground truth, referring episodes, splits, and evaluation metadata will be separately released under CC BY 4.0.

Population	Scenes	Episodes	Episode unit
Benchmark test split	324	3,506	Benchmark query
Controlled fusion/profiles	320	1,762	Object-referring
External core	120	672	Object-referring
Sealed confirmation 1	160	905	Object-referring
Sealed confirmation 2	160	882	Object-referring

Table 3: **Benchmark and evaluation populations.** Controlled fusion and profile analyses share 320 scenes (40 per profile). External core comparisons use a strict subset of 120 scenes (15 per profile). Referring-episode counts denote unique episodes; each is evaluated under both backbones in the controlled and sealed analyses, and under SigLIP in the external comparison. No metric-specific exclusions occur within the controlled or sealed populations shown. Confirmation 1 evaluates the diagnostic oracle-grouped gate; Confirmation 2 provides the primary grouping-free test in Table 1. The sealed query manifests contain 1,726 and 1,715 queries, respectively; these are not the evaluated referring-episode counts.

B Evaluation protocol and statistical analysis

B.1 Frozen implementation

Within each controlled comparison, geometry and semantic outputs are cached by scene, prompt, and configuration digest. All fusion rules and ablations use the same outputs; geometry and readout interventions are evaluated separately.

Track support and validity. In the controlled setting, analytic geometry supplies track identities, canonical positions, and projections, while observation validity remains estimated. This isolates fusion from surface recovery without making the setting independent of ground-truth geometry. In the fully estimated setting, frozen VGGT supplies tracks, positions, projections, and validity cues; ground truth is retained only for evaluation.

For in-frustum track–view pairs, a logistic model maps relative depth margin, depth confidence, and point-map residual to an initial estimate of $P(A_{ti} = 1 \mid \text{geometry})$. Inputs are clipped and standardized by development-set median and IQR; a one-dimensional histogram calibrator then produces η_{ti}. Ground-truth validity A_{ti} supervises fitting only on development scenes. At inference, the frozen geometry-derived η_{ti} is shared across mentions and semantic hypotheses. Off-frustum pairs are masked, not interpreted as observed low responses.

Figure 4 summarizes track support and validity calibration.

Response densities use 48 shared quantile bins with Laplace smoothing. The visible densities f_z are class conditional; the baseline nuisance density c pools all censored observations, whereas h is fitted only to full occlusion and shared between membership labels. Consequently, setting $\rho = 0$ in the type-aware model matches the separately fitted two-component baseline only when $h = c$. The shared h is a modeling approximation motivated by weak measured membership signal under full occlusion; it does not assert exact label independence. All densities and calibration components are fitted on scene-disjoint development scenes. Decision thresholds maximize balanced accuracy on pooled development tracks and are frozen before test evaluation.

SD reliability combines six DAAM-style readouts: two resolutions (16 and 32) at three timesteps (120, 280, and 440). Each readout is standardized by its per-view median and MAD. Let s_{tim} be the standard deviation across the six standardized readouts at projection u_{ti}. Reliability is

$$\beta_{tim} = \mathrm{clip}\big((1 + 0.5 s_{tim})^{-1},\, 0.05,\, 1\big)\,. \tag{5}$$

Equation (5) applies to SD; SigLIP has no corresponding ensemble and uses $\beta_{tim} = 1$ for in-frustum observations. Off-frustum pairs are masked to contribute zero evidence.

Spatial and grouped gates. The grouping-free gate uses the self-inclusive neighborhood in Eq. (4), with radius $0.06D$ and $D = \| \max_t \mathbf{x}_t - \min_t \mathbf{x}_t \|_2 + 10^{-6}$, where extrema are coordinatewise. Thus the scale is the track set's bounding-box diagonal with a numerical safeguard, not its

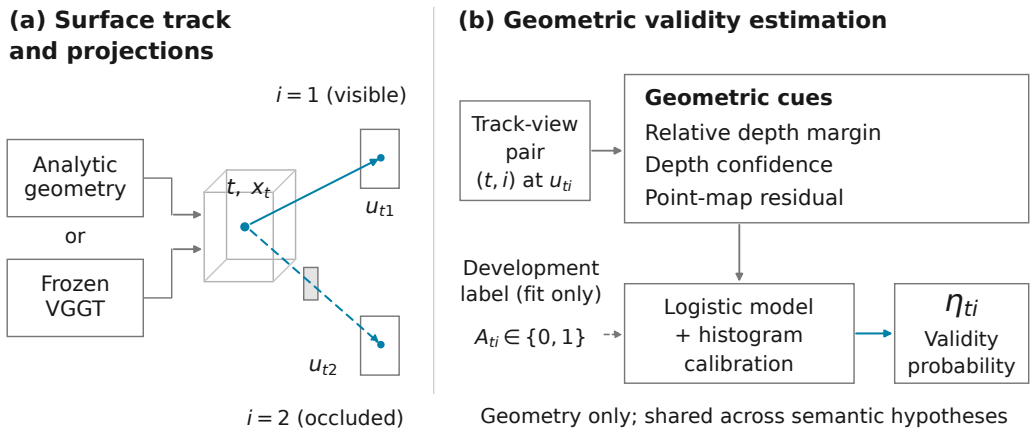

Figure 4: **Track support and estimated observation validity.** (a) A surface track projects to u_{ti} using analytic geometry or frozen VGGT estimates. (b) Relative depth margin, depth confidence, and point-map residual feed a logistic model with histogram calibration to estimate validity probability η_{ti}. Ground-truth visibility A_{ti} provides development-time supervision only (dashed arrow); at inference, the frozen estimator uses geometric cues and is shared across semantic hypotheses. Analytic tracks retain ground-truth support, whereas the fully estimated setting must also reconstruct that support.

maximum pairwise distance. The scored track belongs to its own neighborhood, so the maximum is nonempty and $\rho_{ti}^{\mathrm{sp}} \geq \eta_{ti}$. In the frozen confirmation implementation, validity is set to zero for off-frustum pairs before computing the neighborhood maximum; these pairs also contribute no fusion message. This defines the gate even when some neighbors lie outside the current view.

Controlled confirmation uses analytic positions; the fully estimated setting uses VGGT positions. The neighborhood maximum is a geometric mixture weight, not a calibrated partial-occlusion probability: a visible neighbor can belong to another instance. The diagnostic oracle-grouped gate instead maximizes estimated validity over all tracks in the scored track's ground-truth instance, including the scored track itself. Both gates therefore use estimated visibility. Their sealed comparison tests ranking non-inferiority, not occlusion-type classification accuracy.

Numeric gates for validity quality, censoring gain, backward consensus, grouping, grounding, and transfer are specified before their corresponding test evaluations. We report failed and inconclusive outcomes as well as positive results.

B.2 Fresh sealed confirmations

Repeated development-time evaluation makes the core intervals descriptive. Analysis plans for two disjoint, unseen-seed partitions were recorded before generation. Each partition underwent one original evaluation of its registered endpoints; subsequent post-hoc analyses and reproducibility checks did not alter those records. Confirmation 1 has a documented command clarification made after generation but before evaluation. The original protocols, comparator rules, and realized episode and split manifests are in the anonymous supplementary ZIP (`protocols/`); its `PROVENANCE.md` records the chronology. We present Confirmation 2 first because it supplies the primary grouping-free result in Table 1.

Confirmation 2: primary grouping-free spatial gate. The second partition has 160 scenes and 882 referring episodes per backbone, with no exclusions. Development selection freezes η-weighted LLR as the SD comparator and mean-all as the SigLIP comparator. The type-aware model improves $\mathrm{AUROC}_{\mathrm{occ}}$ by $+0.0503$ $[+0.0453, +0.0552]$ under SD and $+0.0414$ $[+0.0318, +0.0513]$ under SigLIP. Both intervals exclude zero and satisfy the preregistered conjunctive criterion (Table 1). The test-time spatial gate uses no instance-grouping labels, although tracks remain analytic and the response and validity models use labeled development data.

Outcome count	SD 1.5	SigLIP
Spatial gate correct	593	653
Frozen comparator correct	564	507
Both correct	503	474
Spatial gate only correct	90	179
Comparator only correct	61	33
Both incorrect	228	196

Table 4: **Paired target-track selection outcomes.** Each backbone uses 882 episodes across 160 scenes, with no exclusions. The SD comparator is η-weighted LLR; the SigLIP comparator is mean-all. The four joint outcomes partition each population. Net gains are $(90 - 61)/882$ and $(179 - 33)/882$, respectively; confidence intervals resample whole scenes rather than treating episodes as independent.

The preregistered non-inferiority margin against the diagnostic oracle-grouped gate is 0.01 AUROC. Spatial-minus-grouped differences are $-0.0002/-0.0057$ under SD/SigLIP, with one-sided 95% lower bounds of $-0.0024/-0.0074$, both above -0.01. These bounds establish non-inferiority at that margin, without establishing numerical equivalence or calibration of occlusion type.

Target-track selection. On the same 882 episodes, SD selects a target track in $593/882$ cases, versus $564/882$ for its comparator; SigLIP selects correctly in $653/882$ cases, versus $507/882$. The episode-weighted paired gains are $29/882 \approx 0.0328798$ under SD and $146/882 \approx 0.1655329$ under SigLIP, or 3.29 and 16.55 percentage points. Paired scene-bootstrap 95% intervals are $[0.004635, 0.060046]$ and $[0.130289, 0.200673]$, respectively. These intervals were computed in the original sealed run; later verification reproduced their endpoints without changing the analysis. The SD result satisfies the registered directional prediction. SigLIP retains its prespecified descriptive status despite its positive interval. Table 4 gives the exact paired outcomes; candidate and tie rules appear in App. B.5.

Post-hoc partial/full ablation. A later, explicitly post-hoc contrast compares spatial type-aware fusion with the two-component censored mixture on the same 882 episodes and 160 scenes per backbone, with zero exclusions. Both arms retain the cached responses, validity estimates, and their development-fitted densities; the type-aware arm uses the original spatial gate. The two-component baseline uses its pooled censored density c; it is not obtained by simply setting $\rho = 0$ while retaining h. Spatial-minus-two-component gains in $\text{AUROC}_{\text{occ}}$ are $+0.0510$ (95% CI: $[+0.0452, +0.0569]$) under SD and $+0.0608$ ($[+0.0556, +0.0661]$) under SigLIP. The two-component levels are $0.8174/0.8661$, versus $0.8684/0.9269$ for spatial type-aware fusion. These results support the refinement in the grouping-free setting but are not preregistered evidence. Neither the original primary result nor the non-inferiority analysis was altered.

Result records and verification. The supplementary directory `protocols/results/` contains the original Confirmation 2 outputs in `confirm_v2_sd.json` and `confirm_v2_siglip.json`; the corresponding `posthoc_two_component_*.json` files record the labeled ablation, and `r1_verification_*.json` record the hit-count and tie audits. Saved scene-level sums and counts permit recomputation of the paired means and bootstrap intervals. Reproducing the predictions themselves requires the original inference code and episode-level inputs, including renders, cached responses, and geometric cues, which are not included in this metadata package. The protocol files remain unchanged; `protocols/PROVENANCE.md` distinguishes the original evaluations, post-hoc analysis, and later verification.

Descriptive absolute levels for Confirmation 2. Table 1 gives post-hoc per-arm $\text{AUROC}_{\text{occ}}$ levels from the sealed track scores as episode means $\pm$ across-episode standard deviations. They cover the same 882 episodes across 160 scenes per backbone as the registered contrasts. Overall AUROC is 0.891 ± 0.11 under SD and 0.935 ± 0.11 under SigLIP for the spatial-gate model, versus 0.861 ± 0.10 and 0.888 ± 0.16 for the respective frozen comparators. The oracle-grouped reference has $\text{AUROC}_{\text{occ}}$ 0.869 ± 0.14 under SD and 0.933 ± 0.12 under SigLIP. These levels are descrip-

tive; registered inference concerns the paired occluded-tercile contrasts and non-inferiority bounds. Differences and bounds use unrounded values and cannot be recovered from displayed means.

Confirmation 1: supporting diagnostic oracle-grouped gate. The first partition contains 160 scenes and 905 evaluated referring episodes per backbone, with no exclusions. The frozen oracle-grouped model improves $\text{AUROC}_{\text{occ}}$ over its registered non-censored comparator by +0.052 [+0.047, +0.058] under SD and +0.046 [+0.037, +0.055] under SigLIP. Full-occlusion contradiction is approximately 576× and 607× lower than uncensored LLR, respectively. All 16 profile-by-backbone contrasts are positive, with intervals excluding zero. This supports the oracle-grouped model on fresh data; the inference-time removal of grouping labels is tested in Confirmation 2.

B.3 Full-occlusion contradiction

For mention m, let d_{tim} denote the additive evidence contributed by track t in view i. For the mixture rules, this is the reliability-weighted log ratio between the likelihoods at $Z = 1$ and $Z = 0$. Let $\mathcal{C}_m^{\text{full}}$ contain in-frustum track–view pairs with $A_{ti} = 0$ whose physical instance has no visible pixels in view i. Define

$$\text{OCL}_{\text{full}}(m) = \frac{1}{|\mathcal{C}_m^{\text{full}}|} \sum_{(t,i)\in\mathcal{C}_m^{\text{full}}} \max(0, -d_{tim}). \tag{6}$$

Equation (6) is averaged over episodes with nonempty $\mathcal{C}_m^{\text{full}}$ and reported in nats.

OCL is one-sided: it measures negative evidence through fully occluded views, without conditioning on referent membership Z_{tm}. Ground-truth occlusion type defines only the diagnostic set; inference uses estimated validity. Episodes receive equal weight regardless of their number of track–view pairs. Discrimination and calibration, not OCL, assess spurious positive support.

In the shared-h model, $\eta = \rho = 0$ gives exactly zero class evidence. Estimated gates need not reach that endpoint on fully occluded pairs; reduced OCL does not validate the label-independence approximation.

$\text{OCL}_{\text{partial}}$ restricts the diagnostic set to partial occlusion; unsubscripted OCL pools all censored in-frustum pairs. We report a dash for rules such as plain averaging that lack the additive log-evidence decomposition used by this metric.

B.4 Complete controlled-fusion results

Table 5 gives the complete descriptive core comparison, including ranking, operating-point errors, and full-occlusion contradiction. Type-aware rows use the diagnostic oracle-grouped gate; the primary sealed spatial-gate comparison is reported separately in Table 1. Mean-visible and hard-censored LLR retain observations with $\eta_{ti} \geq 0.5$; η-weighted LLR instead scales each uncensored log ratio continuously. The two-component mixture marginalizes validity through Eq. (2), and the type-aware mixture additionally separates partial and full occlusion through Eq. (3).

Hard censoring and mean-visible fusion lose ranking performance despite suppressing invalid observations. Continuous validity weighting improves ranking while limiting contradiction, but treats partial and full occlusion alike. Separating them improves ranking over the two-component mixture, while increasing residual full-occlusion contradiction from $0.001/0.002$ to $0.005/0.011$ nats under SD/SigLIP. Both mixtures remain far below uncensored LLR on this diagnostic. Under SD, removing reliability tempering raises AUROC by 0.015 but doubles OCL_{full} and worsens calibration; SigLIP already uses $\beta = 1$ in-frustum.

B.5 Uncertainty and exclusions

$\text{AUROC}_{\text{occ}}$ is computed within each episode's most-occluded surface tercile and averaged across episodes. Occlusion is summarized by each track's number of truly valid views. Tercile cuts occur only between distinct counts: we choose counts whose cumulative populations are closest to one third and two thirds, breaking equal-distance choices toward the lower count. Tied counts remain together, so tercile sizes need not match. Single-class terciles can still arise from labels, but not from splitting ties. All reported AUROC tables use this tie-preserving construction.

Fusion rule	SD 1.5 responses			SigLIP responses		
	$\text{AUROC}_{\text{occ}} \uparrow$	$\text{FNR}_{\text{occ}} \downarrow$	$\text{OCL}_{\text{full}} \downarrow$	$\text{AUROC}_{\text{occ}} \uparrow$	$\text{FNR}_{\text{occ}} \downarrow$	$\text{OCL}_{\text{full}} \downarrow$
Mean, all views	0.755 ± 0.22	0.331 ± 0.40	–	0.887 ± 0.17	0.158 ± 0.33	–
Mean, visible views	0.633 ± 0.15	0.257 ± 0.30	–	0.683 ± 0.14	0.198 ± 0.30	–
LLR, uncensored	0.787 ± 0.21	0.218 ± 0.35	1.652	0.870 ± 0.18	0.168 ± 0.34	3.539
LLR, hard-censored	0.701 ± 0.14	0.208 ± 0.28	0.001	0.741 ± 0.14	0.208 ± 0.30	0.002
LLR, η-weighted	0.824 ± 0.15	0.286 ± 0.33	0.003	0.874 ± 0.12	0.126 ± 0.27	0.007
Censored mixture, Eq. (2)	0.826 ± 0.15	0.258 ± 0.32	0.001	0.868 ± 0.13	0.111 ± 0.25	0.002
Type-aware mixture, Eq. (3)	0.873 ± 0.14	0.167 ± 0.28	0.005	0.934 ± 0.12	0.093 ± 0.26	0.011
Type-aware, $\beta = 1$	0.888 ± 0.13	0.172 ± 0.29	0.010	n/a ($\beta \equiv 1$ in-frustum)		

Table 5: **Complete controlled-fusion results.** Results use the most-occluded surface tercile of 320 core scenes (40 per profile), fixed analytic tracks, estimated validity, and the same 1,762 object-referring episodes under each backbone, with no exclusions. Type-aware rows use the diagnostic oracle-grouped gate. AUROC and FNR cells report episode means $\pm$ standard deviations; OCL reports mean full-occlusion contradiction in nats. FNR thresholds maximize balanced accuracy on development data and are frozen at test time. These core-population comparisons are descriptive.

Paired uncertainty uses scenes as resampling units because episodes within a scene share renders and reconstruction errors. Unless stated otherwise, intervals use 10^4 scene-level percentile-bootstrap resamples. The same sampled scenes are used for both arms of each paired contrast.

Confirmation 2 bootstrap estimand. The primary contrast, R@1 contrast, non-inferiority analysis, and post-hoc partial/full ablation retain episode weighting when resampling scenes. Let S_g be the sum of paired episode differences in scene g and n_g its number of eligible episodes. A replicate draws G scene indices $g_1, \ldots, g_G$ with replacement and computes $\Delta^* = \sum_{j=1}^{G} S_{g_j} / \sum_{j=1}^{G} n_{g_j}$. All episodes in each sampled scene are retained, including repeated copies when that scene is drawn more than once. This is a ratio of sums, not an equally weighted mean of scene means. Confirmation 2 and its post-hoc ablation use $G = 160$, 10,000 draws, and random seed 0. Two-sided 95% intervals use the 2.5th and 97.5th percentiles. The registered one-sided 95% non-inferiority lower bound uses the 5th percentile: the implementation obtains this as the lower endpoint of a central 90% interval (the helper's two-sided parameter is $\alpha = 0.10$).

Exploratory uncertainty and exclusions. Repeated inspection of core data is not accounted for by nominal intervals. Core contrasts are descriptive even when test labels do not enter model fitting; the sealed partitions provide confirmation.

A coverage diagnostic gives 0.916 coverage at $G = 20$ scenes for nominal 0.950 intervals. Contrasts with $G \leq 25$ are also evaluated using a t-scaled alternative. No exclude-zero conclusion changes, but the percentile intervals are narrower.

AUROC is undefined for single-class occluded terciles. There are no exclusions in the controlled core or either sealed confirmation. Replica occluded-tercile comparisons use 116 of 118 episodes across 38 scenes; ScanNet++ comparisons use 900 of 906 episodes across 220 scenes. Excluded episodes are shared by compared rules. Replica overall-AUROC comparisons use all 118 episodes; interpret counts by endpoint.

R@1-track candidates and ties. R@1-track ranks every track, without a shortlist, and scores a hit if the highest-scoring track belongs to the target instance. All-target or no-target episodes would be excluded; none occurred among the 882 grouping-free confirmation episodes. Exact top ties use first-index order. Neither SD arm nor the SigLIP comparator had ties. The SigLIP spatial arm had 20 tied episodes, but every top tied set had homogeneous membership, so tie order changes no reported R@1 outcome. Ties also occur before sigmoid and are not caused by posterior saturation.

Paired contrast	SD 1.5	SigLIP
Type-aware vs. strongest non-censored	$+0.050$ $[+0.046, +0.054]$	$+0.047$ $[+0.040, +0.054]$
Type-aware vs. two-component mixture	$+0.048$ $[+0.043, +0.053]$	$+0.066$ $[+0.061, +0.070]$
Set $\beta = 1$	$+0.015$ $[+0.012, +0.017]$	n/a
Depth-margin-only η	-0.001 $[-0.002, +0.000]$	-0.001 $[-0.001, -0.000]$

Table 6: **Paired fusion contrasts on the descriptive core population.** Cells give the mean difference in $\text{AUROC}_{\text{occ}}$ and scene-level bootstrap 95% CI, using the same 320 scenes and 1,762 episodes per backbone. The last two rows compare against the default type-aware model. Type-aware rows use the oracle-grouped gate. The non-censored comparators are η-weighted LLR for SD and mean-all for SigLIP. Setting $\beta = 1$ improves SD ranking but worsens calibration and full-occlusion contradiction. SigLIP already uses $\beta = 1$ in-frustum. Signed zeros reflect rounding of the supplied intervals.

C Observation-model ablations and diagnostics

C.1 Occlusion type, reliability, and validity

Table 6 summarizes the descriptive controlled-fusion contrasts using the diagnostic oracle-grouped gate. The post-hoc grouping-free contrast on Confirmation 2 is reported separately in App. B.2; it uses a different population and gate. On the 320-scene population, separating partial and full occlusion improves over the two-component mixture by +0.048/+0.066 $\text{AUROC}_{\text{occ}}$ under SD/SigLIP. Both intervals exclude zero, although these repeatedly inspected core comparisons remain descriptive.

Full-occlusion approximation. Weak measured membership signal under full occlusion motivates shared density h; class-conditional g_z retains partial-occlusion signal. Sharing h discards any residual full-occlusion class information. The zero-evidence endpoint is a model property, not empirical equality of class-conditional full-occlusion distributions. We therefore assess ranking and contradiction separately (App. B.3).

Semantic reliability. Under SD, setting $\beta = 1$ improves $\text{AUROC}_{\text{occ}}$ by 0.015 but raises OCL_{full} from 0.005 to 0.010 nats and worsens calibration. Reliability weighting therefore controls contradiction and calibration rather than explaining the ranking gain. SigLIP already uses $\beta = 1$ in-frustum, so this ablation is inapplicable.

Validity estimator. Replacing the multi-cue estimator with a histogram-calibrated depth-margin cue changes grounding $\text{AUROC}_{\text{occ}}$ by about -0.001 under either backbone. Validity AUROC is 0.934 for the depth-margin cue and 0.952 for the multi-cue estimator. Thus the observed fusion gain does not require the more elaborate predictor. Clipping and robust standardization protect the retained estimator against extreme cues from failed reconstructions.

C.2 Robustness across difficulty profiles

Table 7 partitions the same 320 controlled scenes into eight profiles of 40 scenes each. Each profile uses the diagnostic oracle-grouped gate and the same pooled comparator as the descriptive core contrast in Table 6: η-weighted LLR for SD and mean-all for SigLIP. All 16 intervals exclude zero, and 16 contrasts survive Holm–Bonferroni correction. Descriptive gains span the profiles.

C.3 Cross-view dependence and calibration

Exact posterior log odds require a correctly specified observation model, conditional independence, and an appropriate prior. Our score is a reliability-weighted composite log-likelihood ratio. Mean pairwise correlation across views of the same track is 0.38 under SD and 0.37 under SigLIP. With $n \approx 9.3$ views per track, an exchangeable-correlation approximation gives variance inflation $1+(n-1)\bar{c}$ of roughly four, where $\bar{c}$ is mean correlation. This corresponds to 2.2/2.3 effective independent views.

Difficulty profile	SD 1.5	SigLIP
Baseline	$+0.047$ $[+0.040, +0.055]$	$+0.045$ $[+0.027, +0.065]$
Heavy occlusion	$+0.047$ $[+0.037, +0.058]$	$+0.083$ $[+0.062, +0.107]$
Low overlap	$+0.036$ $[+0.027, +0.045]$	$+0.073$ $[+0.050, +0.097]$
Mixed	$+0.049$ $[+0.036, +0.062]$	$+0.040$ $[+0.021, +0.060]$
Small/thin	$+0.029$ $[+0.019, +0.039]$	$+0.032$ $[+0.018, +0.047]$
Stacked	$+0.057$ $[+0.046, +0.068]$	$+0.017$ $[+0.004, +0.031]$
Touching	$+0.060$ $[+0.049, +0.071]$	$+0.059$ $[+0.039, +0.080]$
Touching + heavy occl.	$+0.068$ $[+0.056, +0.081]$	$+0.039$ $[+0.027, +0.052]$

Table 7: **Type-aware gains by difficulty profile.** Cells give mean $\Delta\text{AUROC}_{\text{occ}}$ and scene-level bootstrap 95% CI over 40 scenes per profile. Results use the oracle-grouped gate and the pooled comparators from the core analysis. All 16 intervals exclude zero and 16 contrasts survive Holm–Bonferroni correction; the weakest lower bound is $+0.004$ (SigLIP, stacked). Reselecting the comparator within a profile or applying the small-G correction leaves every gain positive. These intervals are descriptive.

Dependence is strongest for valid observations (0.56/0.43 under SD/SigLIP), weaker under partial occlusion (0.34/0.36), and near zero under full occlusion (0.09/0.11), where messages are small. It is therefore not confined to the partial-occlusion channel.

Dividing summed composite log-likelihood ratios by a single positive, development-measured variance-inflation factor is a monotone transformation and leaves AUROC unchanged. Occluded-tercile ECE decreases from 0.123 to 0.049 under SigLIP but increases from 0.043 to 0.079 under SD. This intervention shows that a single global correction does not improve calibration for both backbones; its invariance does not establish that dependence has no effect on ranking.

The type-aware mixture has stronger cross-view correlation than the two-component mixture (0.38/0.37 versus 0.27/0.24 under SD/SigLIP), consistent with retaining correlated partial observations. Its ECE is also higher: 0.043 versus 0.039 under SD and 0.123 versus 0.048 under SigLIP. We therefore interpret the score as a composite likelihood and assess calibration separately from discrimination. Dependence-aware calibration is left to future work.

C.4 Spatial extent of the partial channel

For each hidden projection, we measure distance to the nearest visible target pixel and normalize by response spread: approximately 35 pixels for SD and 86 for SigLIP. The mean discrepancy between their evidence–distance profiles falls from 1.20 nats before normalization to 0.26 nats afterward, supporting a readout-dependent spatial scale.

This diagnostic did not improve the gate. Post-hoc sharpening failed to suppress the partial channel as predicted. A development-only image-space gate also failed: preventing reuse of the scored response removed the local signal it sought to exploit. We retain the grouping-free 3D gate with fixed $\gamma = 0.06$ and use response spread only to characterize the mechanism.

D Comparison with learned fusion

Architecture and optimization. The additive MLP uses the same five per-observation inputs as the likelihood model: $(R_{tim}, \eta_{ti}, \rho_{ti}, \beta_{tim}, F_{ti})$, where F_{ti} indicates an in-frustum projection. These inputs are stacked without further preprocessing and passed through a $5 \to 32 \to 32 \to 1$ network with SiLU hidden activations. Its scalar messages are multiplied by F_{ti}, summed over views, and added to a learned scalar prior initialized at the mixture's prior logit. A sigmoid yields the track score.

The network and prior are trained jointly with binary cross-entropy on track membership, using Adam at learning rate 3×10^{-3} for 60 epochs, with one episode per optimizer step and a single shuffling stream. There is no weight decay, early stopping, or checkpoint selection; evaluation uses the final-epoch weights.

Development subsets and fitting failures. We draw five scene-level development subsets at each supervision level $n \in \{1, 2, 5, 10, 25, \text{all}\}$ using one NumPy `default_rng(0)` stream. Response densities, validity calibration, decision thresholds, and learned baselines are refitted on the same subset. At $n = 1$, two of the five subsets leave a response-density stratum empty, so only three yield a fitted mixture and enter the paired comparison. The quoted low-data gains are means of the per-repeat paired mean differences over fittable repeats. They characterize performance conditional on successful fitting, not performance over all sampled one-scene subsets.

Seed sensitivity and comparison rule. The full-supervision sweep fixes the development data and varies 10 training seeds, affecting initialization and episode order. Two executions with seed 0 agree on every reported metric. For each repeat, the MLP is labeled ahead when the scene-bootstrap 95% interval for MLP-minus-mixture $\text{AUROC}_{\text{occ}}$ lies above zero, and behind when it lies below zero. Otherwise the stored label is "tied," meaning statistically unresolved rather than equivalent. Under SigLIP, the MLP is ahead in 5 repeats and unresolved in 5; under SD, it is ahead in 7, unresolved in 2, and behind in 1. These counts do not establish a full-supervision advantage for the structured mixture; we omit a single-seed interval.

Low-data results. With one labeled development scene, the mixture leads by mean gains of +0.032/+0.063 $\text{AUROC}_{\text{occ}}$ under SigLIP/SD, with a positive paired 95% interval in every fittable repeat (three of five subsets). With two scenes, the mean margins are +0.024/+0.045. The MLP becomes competitive after roughly five scenes under SigLIP and ten under SD. The results support a low-data structural advantage where the response densities can be fitted, while exposing a fitting limitation at the smallest supervision level. At $\eta = \rho = 0$, the likelihood model contributes exactly zero class evidence by construction.

E FULLY ESTIMATED GEOMETRY

In the fully estimated setting, VGGT supplies tracks, projections, canonical positions, and the geometric cues used to recompute η_{ti}. The grouping-free gate uses VGGT-estimated positions. Ground-truth geometry is used for evaluation, not for this inference path.

Evaluation-only surface labeling. A seven-degree-of-freedom similarity transform aligns estimated camera centers to ground truth. Estimated tracks are snapped to the nearest surface of the full scene mesh, including room geometry. A track is matched only when the snap distance is strictly below $0.08\,\text{m}$; the matched face supplies its instance label. Room surfaces are legitimate negatives for object-referring mentions. Out-of-bounds tracks are ineligible for matching. Ground-truth validity is recomputed at the snapped point by the exact visibility test. Responses are sampled at each estimated track's own VGGT projections. Unmatched estimates remain in fusion but are excluded from supervised density fitting and all label-based metrics. The estimated arm computes its occluded tercile on the matched tracks' recomputed validity.

Geometry quality and reproducibility. The pipeline covers 0.75 of objects. Reported validity AUROC is 0.903, near the predeclared 0.90 gate. Four query-grid seeds yield 0.9008 ± 0.0018 (range 0.8985–0.9027), with one seed below the gate. Repeated executions at a fixed seed, including runs bypassing the track cache, agree to every reported digit. The observed variation is therefore associated with track initialization. Analytic-track geometric cues reach 0.952 validity AUROC.

Episode pairing does not imply common surface support. On matched estimated tracks, the type-aware mixture reaches 0.931 $\text{AUROC}_{\text{occ}}$ under SigLIP. The separate analytic-versus-estimated contrast pairs episodes, but the analytic arm uses the full ground-truth track population and its occluded tercile, whereas the estimated arm uses matched estimated tracks and its own tercile. Only the estimated arm is restricted to matched recovered support. The reported SigLIP oracle-minus-estimated contrast is -0.002 [-0.015, +0.011], using 404 of 672 episodes across 109 of 120 scenes

for which both arms have defined AUROC. The interval includes zero; it establishes neither equivalence nor geometry error on identical surfaces. The corresponding SD point contrast is -0.158, with estimated tracks scoring higher; this also compares different populations and is not evidence that estimated geometry improves fusion.

Factorial evaluation on reference tracks. The factorial uses a common ground-truth track population. Recovered-support cells (TO) use reference tracks in the intersection of the ground-truth occluded tercile and reached support; these cells share identical reference surfaces. Coverage-aware cells (TE) retain the full reference occluded tercile of the same episodes, including unreached surfaces assigned the evaluator's missing-score floor. TE therefore shares the reference population of the oracle-track cells rather than restricting evaluation to recovered tracks. The populations are nested and answer different ranking and coverage questions.

Score transfer and missing surfaces. For each ground-truth reference track, the evaluator finds the single nearest matched estimated track with the same instance label. Its score is transferred only when the distance is strictly below $0.04D_{\mathrm{gt}}$, where $D_{\mathrm{gt}} = \| \max_t \mathbf{x}_t - \min_t \mathbf{x}_t \|_2 + 10^{-6}$ uses coordinatewise extrema over the scene's ground-truth track set in the world frame, in meters. This is a track-set bounding-box extent, not a scene-mesh diameter. Multiple eligible estimates are resolved by nearest distance, not by maximizing or averaging their scores. This evaluation-only transfer uses ground-truth instance labels; its radius is distinct from the absolute $0.08\,\mathrm{m}$ mesh-matching threshold and the inference gate's $0.06D$ neighborhood, whose point set depends on the inference setting. The same strict $0.04D_{\mathrm{gt}}$ reach criterion defines the reported target-track recall. Reference tracks with no eligible estimate within the radius retain their initial score of $-\infty$. For AUROC, every nonfinite score becomes -10^9. Reached-track sigmoid scores lie in $[0, 1]$, so all unreached reference tracks share a score strictly below every reached track. The same assignment applies to positives and negatives.

Midrank AUROC therefore assigns credit 1 to a reached-positive versus unreached-negative pair, 0 to an unreached-positive versus reached-negative pair, and $1/2$ when both tracks are unreached. Pairs of reached tracks use their transferred-score ordering, with half credit for ties. Thus missing negatives can improve the metric relative to reached positives, and missing targets are not uniformly counted as failures. If all tracks are unreached and both classes are present, AUROC is 0.5. The AUPRC implementation retains the raw $-\infty$ values.

Reported 0.54 $\mathrm{AUROC}_{\mathrm{occ}}$ is the minimum across tested factorial settings, rather than a mean or a selected configuration. Differences among fusion rules are approximately 0.003. Target-track recall is 0.39, computed over all target tracks; it is not restricted to the occluded tercile and does not measure area-weighted surface recovery. An occluded-only recall is not established by this value.

Sampling bias. Image-grid seeding under-samples heavily occluded regions by 2.0–4.7$\times$ relative to mesh sampling. At the one-third visibility threshold, 0.18 of estimated tracks are heavily occluded, versus 0.61 of analytic tracks: a track meets this threshold when visible in at most one third of the views that frame it. Consequently, the recovered population occupies a lower-censoring region of Figure 3. Improving surface discovery is a prerequisite for applying fusion gains to the missing population.

F Transfer regimes and readout dependence

F.1 Interpreting the regime map

Figure 3 summarizes regimes using two development-only quantities: censored-pair fraction and valid-channel overlap $\mathrm{ovl}(f_1, f_0)$. Larger overlap means weaker semantic separation; observed gains concentrate where censoring is frequent and the valid channel remains informative.

Estimated tracks have relatively few censored observations because they under-recover occluded surfaces. Replica exhibits weak semantic separation and partial responses nearly as informative as visible ones. The SAM-crop core readout approaches a high-signal regime with limited remaining headroom, while the fresh synthetic setting resembles the core distribution.

These coordinates are diagnostic, not sufficient to select a fusion rule. The refuted registered ScanNet++ parity prediction also implicates the readout and its fitted likelihood model.

F.2 Replica quality gate and endpoint populations

Replica retains semantic-mesh geometry for evaluation but introduces scanned appearance and reconstruction shift. The predeclared alignment gate leaves 118 test episodes; 116 have a defined $\text{AUROC}_{\text{occ}}$ across 38 scenes. Validity AUROC is 0.903 after excluding 7 of 58 region-scenes whose camera-alignment residual exceeds 0.3 m. Because this gate compares against ground-truth trajectories, the resulting analysis is explicitly conditioned on pose-supervised quality control.

The type-aware ranking contrasts against the strongest non-censored rule are -0.024 [-0.056, +0.006] under SigLIP and -0.034 [-0.070, +0.002] under SD; both intervals include zero. Full-occlusion contradiction nevertheless falls from 1.42 nats under uncensored LLR to 0.11 under the type-aware model. Development densities show partial responses nearly as informative as visible ones, consistent with a clearer effect on contradiction than on ranking.

F.3 ScanNet++ fusion menu and crossed readout

With windowed responses, mean-all reaches 0.762 $\text{AUROC}_{\text{occ}}$, outperforming every likelihood-based rule. The type-aware model reaches 0.726 and uncensored LLR 0.701. The registered contrast against the strongest non-censored rule is negative, -0.036 [-0.051, -0.021].

Within the likelihood family, type-aware fusion improves over uncensored LLR by +0.026, including a gain of +0.009 [+0.005, +0.014] over the two-component mixture. The partial channel is weak: $\text{ovl}(g_1, g_0) = 0.86$, compared with 0.50–0.65 in the other reported regimes. Full-occlusion contradiction is 0.05 nats for the type-aware model versus 0.49 for uncensored LLR, approximately a tenfold reduction.

The valid channel is also weak, $\text{ovl}(f_1, f_0) = 0.75$. This is consistent with limitations of the fitted model for fine-grained window responses in clutter; it does not establish density misspecification. Estimated validity AUROC is 0.766, below the predeclared 0.90 quality gate, providing an additional limitation on interpreting this transfer result.

On the same 900 episodes, mask-crop embeddings reverse the relative ordering: the type-aware model reaches 0.696 $\text{AUROC}_{\text{occ}}$ versus 0.544 for mean-all, exceeding the strongest non-censored rule by +0.021 [+0.016, +0.027]. Both absolute AUROCs are lower than their window-readout counterparts; this reversal is a comparison between fusion rules, not an absolute improvement from changing the readout. This implicates a readout–model interaction without isolating density-estimation error. The partial/full refinement remains beneficial within the tested likelihood family.

G External training-free pipelines

We reimplement OpenScene-style feature averaging (Peng et al., 2023) and MaskClustering-style mask lifting (Yan et al., 2024) using frozen SigLIP features. Table 8 reports the resulting comparisons. Core uses 120 scenes (15 per profile) and 672 object-referring episodes, a strict subset of the controlled 320-scene population. Replica comparisons use 118 episodes overall and 116 with defined occluded-tercile AUROC across 38 scenes.

All methods score the same analytic tracks within each comparison. The two external baselines also receive ground-truth depth and cameras; geometry hyperparameters are selected on development data. The crossed readout/fusion variants use estimated validity and the spatial occlusion-type gate. They do not require ground-truth depth maps for mask backprojection, but this controlled comparison still uses analytic track support and ground-truth camera centers for aligning the VGGT frame. This differs from the fully estimated pipeline in App. E.

Feature averaging. Per-pixel features average normalized SigLIP embeddings from 224×224 windows at stride 96. An observation is retained when it is in-frustum and its projected depth agrees with ground-truth depth within $0.03 + 0.02z$ meters, where z is camera depth. Accepted features are

Method	VisCen3D core		Replica	
	AUROC↑	$\text{AUROC}_{\text{occ}}$ ↑	AUROC↑	$\text{AUROC}_{\text{occ}}$ ↑
Feature averaging (OpenScene-style)	0.898 ± 0.11	0.871 ± 0.13	0.722 ± 0.25	0.705 ± 0.28
Mask lifting (MaskClustering-style)†	0.961 ± 0.07	0.951 ± 0.08	0.710 ± 0.23	0.724 ± 0.23
Mean fusion, all views	0.882 ± 0.17	0.883 ± 0.17	0.726 ± 0.25	0.734 ± 0.26
LLR, η-weighted	0.893 ± 0.11	0.876 ± 0.12	0.713 ± 0.25	0.680 ± 0.28
Type-aware mixture, window readout	0.935 ± 0.11	0.932 ± 0.12	0.727 ± 0.26	0.710 ± 0.29
SAM-crop readout, mean fusion	0.827 ± 0.15	0.821 ± 0.15	0.649 ± 0.28	0.635 ± 0.28
SAM-crop readout, type-aware mixture	0.977 ± 0.05	0.963 ± 0.07	0.712 ± 0.25	0.610 ± 0.31

Table 8: **External pipelines and crossed readout/fusion comparisons.** All rows use SigLIP and common analytic track support. Cells report episode means $\pm$ standard deviations, not confidence intervals. Core uses 672 episodes across 120 scenes for both endpoints. Replica uses 118 episodes for overall AUROC and 116 for $\text{AUROC}_{\text{occ}}$, across 38 scenes. External baselines receive ground-truth depth and cameras; crossed variants use estimated validity and the spatial gate, while retaining the controlled analytic support and camera-frame alignment. †Development selection reaches the no-merge limit.

averaged, ℓ_2-normalized, and scored by cosine similarity to the text embedding. If no view passes, the method uses all in-frustum observations.

Mask lifting. SAM ViT-B masks (Kirillov et al., 2023) are generated on a 16×16 point grid and backprojected with ground-truth depth. After area filtering, each mask contributes at most 600 3D points. Voxelized proposals are merged when voxel-containment overlap exceeds τ, then scored by SigLIP using their five largest padded crops. Tracks inherit proposal scores under a 3^3 voxel-neighborhood vote. Voxel size and τ are selected on development data. On both partitions, selection reaches the no-merge limit, yielding per-mask retrieval. Tracks assigned to no proposal receive the scene floor, defined as the minimum assigned score minus 10^{-3}. If no track is assigned, the fallback score is 0 for all tracks. Unassigned tracks remain in evaluation and tie below every assigned track when assignments exist; AUROC uses midranks for ties. The unassigned fractions are 0.5% on core and 13.2% on Replica.

Crossing readout and fusion. Each mask's crop score is painted onto its image pixels, averaging overlaps. Uncovered pixels, including those in views with no masks, receive the minimum over all crop scores in the scene minus 10^{-3}. This precedes scene-pooled CDF normalization and track sampling. The maps then enter mean fusion or the type-aware mixture. The floor applies to pixel responses, not final track scores: likelihood mapping and fusion can change the ranking of tracks sampling uncovered pixels.

On core, mean fusion with crop scores reaches 0.827 overall AUROC, whereas type-aware fusion reaches 0.977/0.963 overall/occluded AUROC. The latter exceeds mask retrieval by +0.016 overall and +0.012 [+0.007, +0.018] on occluded surfaces. Thus crop readouts and censored fusion can work together on core.

The combination does not transfer uniformly. On Replica, the crop-readout mixture reaches 0.610 $\text{AUROC}_{\text{occ}}$ and trails mask lifting by -0.114 [-0.163, -0.064]. The window-readout mixture instead has contrasts with feature averaging and mask lifting whose intervals include zero: +0.005 [-0.041, +0.052] and -0.014 [-0.069, +0.037], respectively.

System	Overall	Ambig.	Plain	Attr.	Null
Semantic unaries only	54.4	34.8	80.0	80.0	10.0
Full factor graph	67.6	60.9	95.0	80.0	10.0
Deterministic predicates	67.6	60.9	95.0	80.0	10.0
Null state removed	66.2	60.9	95.0	80.0	0.0
Oracle entities	66.2	56.5	95.0	80.0	10.0

Table 9: **Downstream grounding on scenes containing twins.** Accuracy (%) on 68 queries across five held-out test scenes. Category counts are 23 ambiguous relational (Ambig.), 20 unambiguous relational (Plain), 15 attribute (Attr.), and 10 null queries. Categories are disjoint and exhaustive; all rows use identical queries without exclusions. Overall accuracy pools queries ($46/68$ for the full graph). Indented rows modify the full graph. Relations improve the tested relational cases, while null-target accuracy remains low. The separate oracle-entity predicate diagnostic is not a query-level result.

H Downstream demonstrations and qualitative results

The observation model scores physical surface tracks. The following downstream illustrations are separate from the primary grouping-free confirmation in Table 1; they do not evaluate a complete deployable grounding system.

Backward consensus. Messages from all but one view are projected into the held-out view using validity-gated, support-normalized Gaussian kernels. This leave-one-view-out construction does not reuse the target view's message. In the principal backward-consensus analysis, per-view mIoU increases by +5.5 points and cross-view disagreement decreases by 32.5%, without degrading the easiest views. A fresh partition gains +2.9 mIoU points (CI [+1.1, +4.5]); an intermediate partition gains 1.5 points with an interval crossing zero. The refinement is modest and depends on the regime.

Track grouping. Tracks are grouped using frozen SAM mask co-membership, 3D proximity, and semantic agreement, with an evidence-based veto against merging. No clearly separated objects are merged in the evaluated cases. Identical touching instances remain difficult: their merge rate falls from 50% with pairwise mask evidence to 10% with a mask–track hypergraph. The latter reaches the predeclared gate boundary, so this result does not establish reliable separation of touching twins.

Relational grounding. A factor-graph executor combines entity scores, query-conditioned geometric predicates, explicit reference frames, and a calibrated null state. Eight touching-pair scenes are split into three development and five test scenes. Table 9 evaluates the same 68 queries from the five test scenes in every row, with no exclusions. Its categories are disjoint and exhaustive: 23 ambiguous relational, 20 plain relational, 15 attribute, and 10 null queries. Overall accuracy pools queries rather than averaging scenes or categories. The full graph answers $46/68$ queries correctly, versus $37/68$ for semantic unaries; the null-removed and oracle-entity variants each answer $45/68$ correctly.

Pooling the two relational categories gives $33/43$ correct for the full graph versus $24/43$ for semantic unaries, a +20.9-point gain. The 99.9% directional-predicate accuracy with oracle entities is a separate diagnostic over $7{,}110$ sampled predicate evaluations, not query-level grounding accuracy. Oracle entities do not improve the table's overall grounding accuracy, and removing the null state lowers null accuracy from $1/10$ to $0/10$. Inspected null failures involve semantic distractor responses, but this does not isolate the readout as their sole cause.

Figure 5 shows selected successes and failures. These examples illustrate local behavior and do not supersede the aggregate ScanNet++ contrast, which favors mean-all with windowed responses.

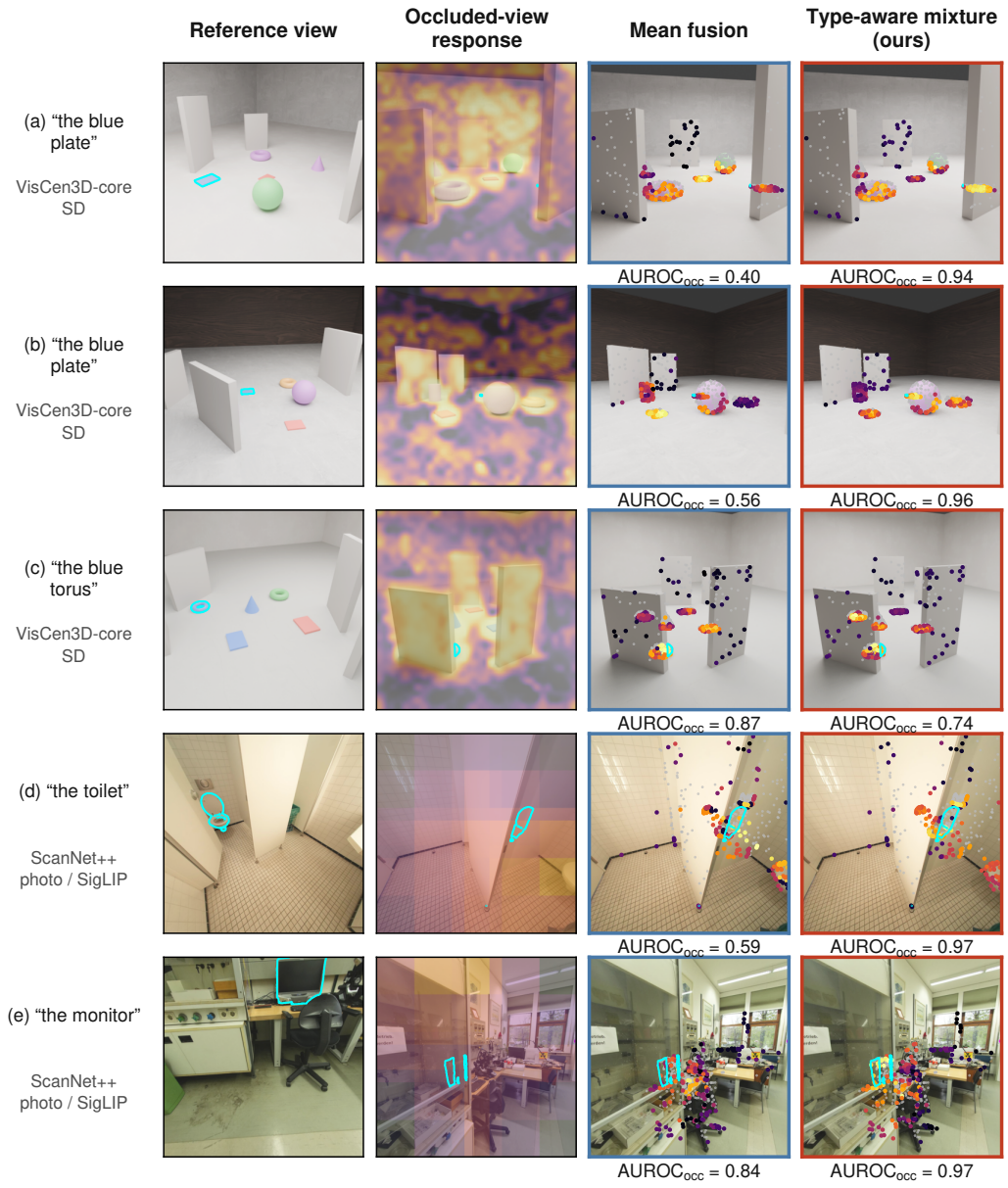

Figure 5: **Qualitative grounding under occlusion.** Each row shows a reference view with the target outlined in cyan, an occluded view with its frozen semantic response, and track rankings from mean fusion and the type-aware model. Colored dots mark tracks in the most-occluded surface tercile, with color encoding within-episode rank; gray dots provide visible-track context. Values below the fusion panels report AUROC$_{occ}$. Rows (a)–(b) show VisCen3D-core successes where the type-aware model better preserves hidden-target evidence. Row (c) shows a failure on a thin, barely observed target, with elevated distractor rankings. Rows (d)–(e) show selected ScanNet++ successes consistent with suppressing invalid negative evidence; these examples do not represent the aggregate ScanNet++ result, where mean fusion performs better with windowed responses.