BEYOND ARGMAX: A MECHANISTIC STUDY OF SEMANTIC RETENTION IN FROZEN FOUNDATION-MODEL COMPOSITION FOR GENERALIZED FEW-SHOT 3D SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical classifier-combination work distinguishes score-level fusion from hard decision voting. We revisit this distinction in a modern regime where independently pretrained, frozen foundation models are composed at inference time for generalized few-shot 3D segmentation. Our contribution is not a new sum rule; we ask a mechanistic question: *how much useful semantic information is lost when heterogeneous sources are collapsed to one class before interaction?* We construct a same-input top-k retention intervention that freezes model weights, point clouds, masks, geometry, class vocabularies, scene lists, and fusion rules while varying only the semantic support retained before fusion. On 156 held-out ScanNet200 scenes, full retention reaches 34.87 harmonic-mean (HM) IoU versus 28.47 for the frozen top-1 control ($+6.40$, 95% paired scene-bootstrap CI $[+5.24, +7.64]$). The effect independently replicates on all 50 ScanNet++ validation scenes (26.50 vs. 23.02 HM; $+3.48$, CI $[+1.64, +5.93]$). Most of the gain appears immediately after top-1, while moderate top-k can slightly outperform the unrestricted tail. The conclusion survives a GroundingDINO–SAM2.1 sparse-source replacement, alternative fusion operators, tie-policy changes, and source-weight variation. Calibration diagnostics show severe but opposite raw miscalibration of the dense and sparse sources, yet calibration alone does not eliminate the retention advantage. We therefore identify *premature semantic collapse* as a repeatable information bottleneck in heterogeneous frozen-model composition.

1 INTRODUCTION

Frozen foundation models are increasingly composed rather than retrained end-to-end. In 3D perception, a system may combine a language-aligned point-cloud model, promptable image segmentation, camera geometry, and lightweight inference-time logic. This creates a simple but consequential design choice: *when should semantic uncertainty be collapsed into a discrete label?*

The abstract distinction is classical. Multimodal classification frameworks distinguish decision-level voting from score- or probability-level combination; sum, product, and max rules have been studied for decades (Kittler et al., 1998; Kuncheva, 2004; Baltrušaitis et al., 2018). We therefore do **not** claim that averaging distributions is novel. Instead, we study whether early collapse is a practically important failure mode in a different regime: heterogeneous, independently pretrained, frozen foundation-model sources with partial spatial overlap, cross-view lifting, and generalized few-shot base–novel evaluation.

A hard pipeline applies

$$\text{source distributions} \rightarrow \arg\max \rightarrow \text{selection/voting}, \tag{1}$$

whereas a retention-preserving pipeline delays the irreversible step,

$$\text{source distributions} \rightarrow \text{cross-source interaction} \rightarrow \arg\max . \tag{2}$$

If a dense 3D model and a sparse 2D-derived source both retain several plausible categories, a second-ranked class from one source can reinforce compatible evidence from the other. Once either source is reduced to top-1, no later calibrator or selector can recover that alternative.

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.

We isolate this mechanism using frozen RegionPLC (Yang et al., 2024) as the dense source and cross-view SAM3 (Carion et al., 2026) concept evidence as the primary sparse source. A matched top-k intervention keeps upstream evidence fixed and changes only how much semantic information survives before fusion. The resulting curves reveal a consistent pattern on ScanNet200 (Dai et al., 2017; Rozenberszki et al., 2022) and ScanNet++ (Yeshwanth et al., 2023): the largest improvement occurs immediately after relaxing top-1, followed by saturation. A GroundingDINO (Liu et al., 2024)–SAM2.1 (Ravi et al., 2024) replacement diagnostic reproduces the effect with a different sparse stack.

Contributions.

1. **Mechanistic diagnosis.** We explicitly connect frozen 3D foundation-model composition to classical decision- versus score-level fusion and isolate *premature semantic collapse* as the object of study.
2. **Matched information intervention.** A same-input top-k ladder fixes sources, masks, geometry, vocabularies, scene sets, and fusion while varying only retained semantic support.
3. **Cross-dataset evidence.** Full retention improves HM over matched top-1 by 6.40 points on ScanNet200 and 3.48 on ScanNet++, with paired scene-bootstrap intervals excluding zero.
4. **Robustness across source stacks and fusion choices.** The effect persists under GroundingDINO–SAM2.1, tie-policy changes, weight sweeps, and max/geometric pooling; we also report calibration and computational diagnostics.

2 Related Work

Decision- and score-level combination. Kittler et al. (1998) and Kuncheva (2004) formalized sum, product, max, and voting rules, establishing the framework for combining classifier outputs. Retaining continuous distributions is likewise central to product-of-experts formulations (Hinton, 2002) and deep ensemble uncertainty estimation (Lakshminarayanan et al., 2017). Beyond classical ensembling, modern foundation models are frequently composed via parameter merging (Wortsman et al., 2022) or distribution-preserving output combination; DeePEn (Huang et al., 2024), for example, fuses heterogeneous language-model distributions after mapping them to a shared relative space. Our question is narrower: whether semantic collapse before spatially heterogeneous frozen sources interact creates an information bottleneck in 3D perception, and how much support must survive.

Generalized few-shot 3D segmentation. Early few-shot point cloud segmentation focused on prototype alignment and metric learning (Zhao et al., 2021; Chen et al., 2021), while non-parametric architectures (Zhang et al., 2023) demonstrated the utility of training-free point cloud analysis. Generalized few-shot segmentation (GFS-PCS) evaluates both base-class retention and novel-class adaptation via their harmonic mean (Xian et al., 2019). Recent methods adapt representations through vision–language pretraining: GFS-VL (An et al., 2025) combines few-shot samples with pseudo-labels, while HOP3D (Zhao et al., 2026) suppresses base–novel interference through hierarchical orthogonal prototypes. These trained methods provide contextual performance references, but are not causal baselines for our intervention because they update representations. Our primary comparisons keep upstream evidence frozen.

Frozen 2D–3D composition. Transferring representations from 2D vision foundation models (Radford et al., 2021; Kirillov et al., 2023) into 3D geometric spaces is standard in open-vocabulary perception. Landmark pipelines lift multi-view 2D features to points or radiance fields via geometric projection (Peng et al., 2023; Ding et al., 2023; Jatavallabhula et al., 2023; Kerr et al., 2023). In training-free 3D segmentation, RegionPLC (Yang et al., 2024) provides language-aligned point representations, while OV-SAM3D (Tai et al., 2024) and MaskClustering (Yan et al., 2024) aggregate multi-view 2D instance masks into 3D consensus structures. Cross-modal distillation methods such as FtD++ (Wu et al., 2025) exploit complementary 2D/3D modalities. We begin after semantic evidence is produced and ask what distribution support must remain available at combination time.

Confidence calibration and reliability. Deep neural networks frequently exhibit overconfidence and systematic miscalibration (Guo et al., 2017; Minderer et al., 2021). Post-hoc techniques such

as histogram binning (Zadrozny & Elkan, 2001) adjust confidence estimates to reflect empirical correctness, assessed via Expected Calibration Error (ECE) and Brier scores (Brier, 1950). Selective classification frameworks (Geifman & El-Yaniv, 2017) similarly use calibrated confidence to determine prediction routing or rejection. We evaluate whether confidence miscalibration accounts for the gap between hard and soft composition, testing whether calibrating frozen sources can bridge the deficit caused by premature collapse.

3 Semantic-Retention Intervention

Let $u \in \mathcal{U}$ be a 3D point and $\mathcal{C}$ the runtime vocabulary. Dense (d) and sparse (s) frozen sources provide normalized semantic vectors $\boldsymbol{\pi}_s(u) \in \mathbb{R}^{|\mathcal{C}|}_{\geq 0}$ when available. A distribution-level rule forms

$$\boldsymbol{\pi}_{\text{fuse}}(u) = \frac{\sum_s a_s(u)\beta_s(u)\boldsymbol{\pi}_s(u)}{\sum_s a_s(u)\beta_s(u)}, \qquad \widehat{y}(u) = \arg\max_c \pi_{\text{fuse},c}(u), \tag{3}$$

with dense fallback when sparse evidence is absent.

Why top-1 is restrictive. The map $\boldsymbol{\pi} \mapsto \arg\max_c \pi_c$ is many-to-one. The following elementary proposition motivates the intervention rather than claiming new fusion theory.

Proposition 1. *No rule observing only per-source argmax labels can reproduce equal probability fusion for all source distributions.*

Proof. For two classes, consider $([0.90, 0.10], [0.49, 0.51])$ and $([0.51, 0.49], [0.10, 0.90])$. Both expose the same hard tuple $(1, 2)$, yet equal probability averaging predicts class 1 for the first pair and class 2 for the second. A rule receiving only $(1, 2)$ must return the same answer for both. □

Matched top-k retention. Let $\mathcal{T}_k(\boldsymbol{\pi})$ keep the k largest entries and renormalize. The primary intervention is

$$\boldsymbol{\pi}^{(k)}_{\text{fuse}}(u) = \tfrac{1}{2}\mathcal{T}_k(\boldsymbol{\pi}_d(u)) + \tfrac{1}{2}\mathcal{T}_k(\boldsymbol{\pi}_s(u)), \quad k \in \{1, 2, 5, 10, 20, \text{all}\}. \tag{4}$$

Thus only semantic support changes; the evidence, geometry, and 0.5/0.5 rule do not.

Dense RegionPLC. RegionPLC (Yang et al., 2024), operating on sparse 3D convolutional representations (Choy et al., 2019), produces class logits over the vocabulary. Logits are converted to probabilities before argmax. ScanNet200 uses a 200-way source space and ScanNet++ a matched 30-class space. Probabilities are cached as float16 and converted to float32 only in streaming evaluation chunks.

Sparse SAM3 cross-view evidence. Each canonical class name is prompted independently and verbatim into frozen SAM3 (Carion et al., 2026); we use confidence threshold 0.3 and every 20th captured frame. Multiple instances of the same concept in one frame are unioned. A 3D point is projected with the color-camera intrinsics and inverse camera pose. With $m_{v,c}(u)$ indicating membership in concept c's frame-v union mask and $\nu_v(u)$ valid in-frustum visibility,

$$q_u(c) = \frac{\sum_v m_{v,c}(u)}{\max(1, \sum_v \nu_v(u))}, \qquad \pi_{s,c}(u) = \frac{q_u(c)}{\sum_j q_u(j)}. \tag{5}$$

Only nonzero $(u, c, q_u(c))$ entries are stored. The main matched study uses no depth z-buffer. On ScanNet++, sparse evidence covers the 18 novel evaluation classes and has zero mass on the 12 base classes; this asymmetry is explicitly reflected in the reported base–novel trade-off.

4 Experimental Protocol

ScanNet200. All 312 validation scenes from ScanNet200 (Rozenberszki et al., 2022), built upon the ScanNet benchmark (Dai et al., 2017), are ordered once and split by parity into 156 fit and 156 evaluation scenes. Calibration/reliability components use labels only from the fit half; all ScanNet200 results below use the same held-out 156 scenes. We report base mIoU, novel mIoU, and their harmonic mean (HM) (Xian et al., 2019). Headline differences use paired scene bootstrap confidence intervals (5,000 replicates) (Efron & Tibshirani, 1994).

Table 1: Held-out ScanNet200 (156 scenes). All methods reuse frozen upstream evidence.

Method	Base	Novel	HM
Dense RegionPLC	43.60	19.06	26.53
Frozen hard selector	41.84	20.49	27.51
Top-1 one-hot (frozen)	45.13	20.80	28.47
Linear full distribution	44.43	28.70	**34.87**
Calibrated distribution	45.44	27.82	34.51
Global reliability	45.70	27.56	34.38
Contextual distribution	44.06	20.31	27.80

Table 2: Matched semantic-retention ladders. Confidence intervals are for Full minus Top-1.

	ScanNet200			ScanNet++		
Retention	Base	Novel	HM	Base	Novel	HM
Top-1	45.13	20.80	28.47	32.12	17.94	23.02
Top-2	45.48	24.49	31.83	27.52	25.85	26.66
Top-5	47.42	27.08	34.48	27.37	26.39	26.87
Top-10	47.33	28.02	35.20	27.31	26.54	26.92
Top-20	46.95	28.53	**35.49**	27.27	26.57	**26.92**
Full	44.43	28.70	34.87	27.21	25.83	26.50
Full − Top-1	$+6.40$ $[+5.24, +7.64]$			$+3.48$ $[+1.64, +5.93]$		

ScanNet++. The parameter-free retention ladder is independently evaluated on all 50 validation scenes in a 30-class space (12 base/18 novel), with aligned geometry, undistorted DSLR imagery, and corresponding camera transforms. No fit split is required for the fixed retention rule.

Source replacement. For GroundingDINO–SAM2.1, five deterministically selected ScanNet200 scenes and ten deterministic frames per scene are used. Three GroundingDINO prompt chunks cover all 200 concepts; proposal threshold is 0.35 and near-identical boxes are clustered at IoU ≥ 0.90. SAM2.1 Hiera-L produces one mask per spatial cluster. Across 50 frames, 511 clusters are shared by all retention conditions. Hard IDs are the exact argmax collapse of the stored 200-way soft support: zero argmax, concept, and hard-mask-union mismatches are observed. Here RegionPLC remains full-distribution and only the replacement sparse semantics are truncated.

5 Results

5.1 ScanNet200: Retention Matters More Than Weighting Complexity

Table 1 shows the frozen held-out result. Dense RegionPLC reaches 26.53 HM; the frozen hard selector reaches 27.51; and the original matched top-1 control reaches 28.47. Uniform full-distribution averaging reaches 34.87 HM, improving over top-1 by 6.40 points (95% CI $[+5.24, +7.64]$) and over the frozen hard selector by 7.37 points (CI $[+6.29, +8.44]$). Calibrated/global weighting remain close to the simple average, while contextual weighting falls near the hard-selector regime. This negative result is informative: more elaborate weighting cannot reconstruct semantic alternatives already destroyed by argmax.

5.2 Where the Information Gain Appears

The matched ladder (Table 2) exposes the mechanism. On ScanNet200, merely retaining the second hypothesis raises HM from 28.47 to 31.83. Performance increases to 35.49 at top-20 before the unrestricted tail settles at 34.87. ScanNet++ shows the same geometry: 23.02 at top-1 jumps to 26.66 at top-2 and saturates near 26.9. Thus the supported claim is not "keep every probability" but "do not collapse to one class before interaction."

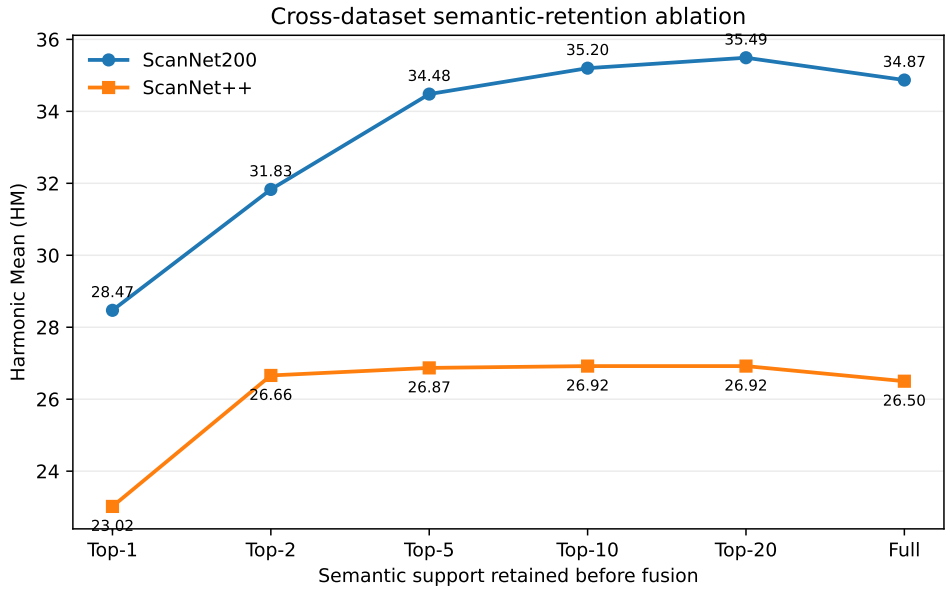

Figure 1: The same retention geometry appears on ScanNet200 and ScanNet++: a large Top-1→Top-2 gain followed by saturation.

Table 3: GroundingDINO–SAM2.1 source-replacement diagnostic (five fixed ScanNet200 scenes). Only sparse semantic retention changes.

Retention	Base	Novel	HM	ΔHM
Top-1	21.14	11.35	14.77	0.00
Top-2	22.62	11.91	15.61	+0.84
Top-5	24.20	12.54	16.52	+1.75
Top-20	26.22	13.80	18.08	+3.31
Full	27.56	14.21	**18.75**	+3.98

5.3 Source-Stack and Fusion-Operator Robustness

Replacing SAM3 with GroundingDINO–SAM2.1 preserves the phenomenon despite weaker absolute sparse evidence. On the five-scene controlled diagnostic, sparse semantic retention increases HM monotonically from 14.77 (Top-1) to 18.75 (Full), a $+3.98$ diagnostic gain with scene-bootstrap interval $[+1.11, +6.54]$. Dense-only is 20.70 HM, so this experiment supports *source-independence of the retention effect*, not superiority of the replacement sparse predictor.

The result also survives classical distribution-level operators on the full held-out ScanNet200 set (Table 4). Using the conservative dense-favouring Top-1 control (26.53 HM), max pooling reaches 30.41, linear pooling 34.87, and geometric/log pooling (Product of Experts) (Hinton, 2002) reaches 36.16. We retain linear averaging as the frozen headline rule because operator tests were performed after the primary study; PoE is reported only as robustness evidence. Its HM varies by just 0.16 across $\epsilon \in [10^{-12}, 10^{-4}]$.

Tie and weight robustness. At equal-weight one-hot fusion, dense/sparse Top-1 disagreement creates an exact tie on 67.93% of sparse-covered valid points. The frozen evaluator's lowest-index tie rule gives the *strongest* tested Top-1 baseline (28.47 HM); dense- and sparse-favouring alternatives give 26.53 and 20.02 HM, respectively. Full retention therefore improves HM by at least 6.40 points under every deterministic tie policy tested. Likewise, full-distribution HM stays in a narrow 34.17–34.87 range as the sparse weight varies from 0.3 to 0.7, whereas one-hot weighting degenerates into a hard source switch around $w_s = 0.5$. Full tables are in Appendix A.

Table 4: Selected robustness diagnostics on held-out ScanNet200.

Diagnostic	Base	Novel	HM
Top-1, dense-favouring	43.60	19.06	26.53
Max full distribution	37.59	25.53	30.41
Linear full distribution	44.43	28.70	34.87
Geometric/log full	47.07	29.35	**36.16**

Table 5: Held-out source calibration. Accuracy, mean confidence, and ECE are percentages; calibration is fitted only on the 156-scene fit half.

Source	State	Acc.	Conf.	ECE ↓	Brier ↓
Dense	raw	61.80	40.88	29.20	0.3117
Dense	calibrated	61.80	62.50	1.54	0.2146
Sparse	raw	38.02	81.31	43.29	0.4187
Sparse	calibrated	38.02	37.74	0.69	0.2190

Calibration. On the held-out 156 scenes, dense RegionPLC is strongly under-confident (61.80% empirical accuracy vs. 40.88% mean confidence; ECE 29.20% (Guo et al., 2017)), while sparse SAM3 is strongly over-confident (38.02% accuracy vs. 81.31% confidence; ECE 43.29%). Histogram calibration (Zadrozny & Elkan, 2001) fitted only on the 156 fit scenes reduces ECE to 1.54% and 0.69%, and Brier score (Brier, 1950) from 0.3117 to 0.2146 (dense) and 0.4187 to 0.2190 (sparse). Yet calibrated distribution fusion remains slightly below simple linear fusion (34.51 vs. 34.87 HM), separating confidence miscalibration from semantic-information loss.

5.4 Context Against Trained GFS-PCS Systems

Our frozen mechanism study is not intended to replace trained few-shot adaptation. Table 6 gives absolute context using 1-shot numbers reported by the original trained methods. These are not causal baselines: GFS-VL/HOP3D update task-specific representations, whereas our intervention performs no parameter updates and holds upstream evidence fixed.

6 Mechanistic Failure Analysis

Quantitative gains are not uniform. Positive qualitative examples were selected only after the experiment was frozen using deterministic criteria ($\Delta\mathrm{HM} > 0$, $\Delta\mathrm{Novel} > 0$, and positive net point corrections), avoiding visual cherry-picking. The mechanism map in Figure 2 compares matched Top-1 with Full at point level: improvements are spatially localized rather than a global relabeling. We also retain `scene0490_00` as an explicit failure: Top-1 reaches 31.66 HM while Full falls to 19.41 (-12.25), with $-12{,}015$ net point corrections. This supports adaptive retention or selective classification (Geifman & El-Yaniv, 2017) rather than indiscriminate preservation of the entire tail.

7 Discussion and Limitations

What is new. The classical score-versus-decision distinction is not new; our contribution is the controlled diagnosis of its magnitude and structure in heterogeneous frozen foundation-model composition. The repeated Top-1→Top-2 jump across ScanNet200 and ScanNet++ is informative: most useful uncertainty is concentrated in a compact candidate set. The GroundingDINO–SAM2.1 replacement confirms that this finding is not specific to SAM3.

Representation before selector complexity. Calibration and global weighting improve confidence handling but do not outperform simple retained-distribution averaging. Once alternatives are destroyed, a more sophisticated selector cannot reconstruct them. Conversely, retained distributions permit genuinely graded weighting; one-hot evidence turns source weights into near-binary source choice.

Table 6: Contextual 1-shot HM from original trained-method reports versus the frozen linear-retention result. Protocol/training differences make this contextual rather than a matched comparison.

Method	ScanNet200 HM	ScanNet++ HM
GFS-VL (An et al., 2025)	40.92	29.47
HOP3D (Zhao et al., 2026)	43.42	29.32
Frozen linear retention (ours)	34.87	26.50

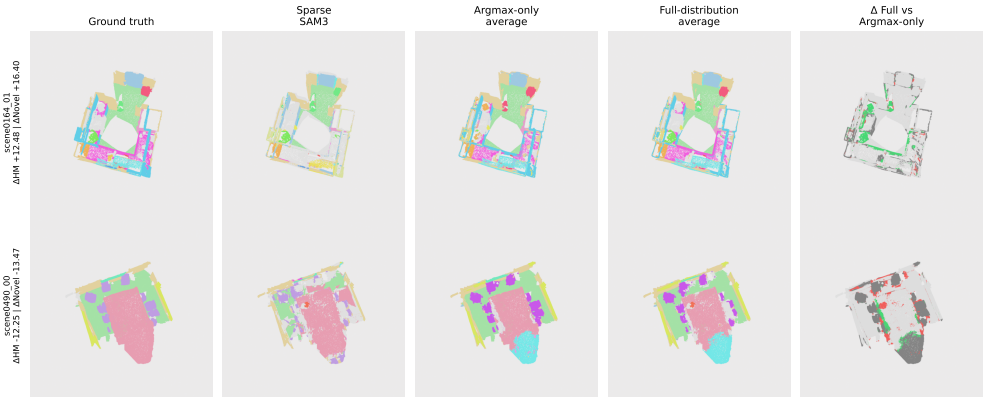

Figure 2: Matched point-level mechanism visualization (file supplied with the submission): beneficial Full-vs.-Top-1 changes, harmful changes, and points unchanged in correctness. Extended positive/failure examples are provided in the appendix/supplement.

Failure modes. Full tails are not always beneficial. Moderate top-k slightly outperforms full retention on both main datasets, and an explicit ScanNet200 failure scene loses 12.25 HM under Full relative to Top-1. This motivates future adaptive k or entropy/agreement-aware gating rather than the stronger claim that unrestricted tails should always be retained. ScanNet++ also has asymmetric sparse support: the sparse source covers novel classes only, increasing novel mIoU while reducing base mIoU under stronger retention.

Efficiency. The full 312-scene ScanNet200 cache contains 49.54M points and occupies 13.31 GiB compressed; raw float16 200-way dense probabilities require 18.46 GiB and sparse COO evidence 2.93 GiB. Streaming fusion in 5,000-point chunks has measured peak RSS 719.5 MiB. On CPU, fusing all 25.02M held-out points takes 12.0 s (2.08M points/s); compressed-cache loading, not fusion, dominates wall-clock time. The observed top-k saturation provides a direct path toward smaller candidate-list storage.

Scope. The study uses particular frozen source stacks and indoor 3D datasets; it does not establish universal superiority of probability fusion, nor does it argue that training or adaptation is unnecessary. The five-scene GroundingDINO–SAM2.1 experiment is a controlled diagnostic rather than a third benchmark. Further source families, outdoor scenes, and adaptive retention policies are natural extensions.

8 Conclusion

We studied semantic collapse as an information bottleneck in frozen foundation-model composition rather than proposing probability averaging as a new fusion rule. Across ScanNet200 and ScanNet++, matched top-k interventions show that collapsing to one class before interaction discards useful alternatives, with most of the recoverable information appearing immediately beyond top-1. The effect survives source replacement, fusion-operator changes, tie-policy choices, and source-weight variation. Calibration corrects severe confidence mismatch but does not recover discarded class hypotheses. The resulting design principle is simple: *when heterogeneous frozen models must interact,*

preserve a compact set of plausible semantic alternatives until after cross-source evidence has been combined.

Reproducibility Statement

The main text specifies the frozen sources, class spaces, frame sampling, projection and sparse-evidence construction, retention operator, scene splits, metrics, and statistical protocol. Appendix A provides complete tie, weight, calibration, efficiency, and numerical-stability diagnostics. Code and frozen evaluation caches are planned for anonymized release with the submission to reproduce the same-input comparisons.

AI Assistance Statement

Generative AI tools (Large Language Models) were used solely to assist with manuscript drafting, language polishing, LaTeX typesetting, and literature discovery/retrieval. All core research hypotheses, theoretical formulations, experimental designs, quantitative benchmarks, and empirical findings were conceived, conducted, and evaluated entirely by the authors on the stated models and datasets. All AI-assisted text, citations, and mathematical notation were rigorously verified by the authors against source logs and primary literature. The authors take full responsibility for the entire contents and integrity of this submission.

References

Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. Generalized few-shot 3D point cloud segmentation with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16997–17007, 2025.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2018.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Somnath Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Runnan Chen, Xinge Chen, and Bingbing Zhou. PAP-FSS: Point adversarial prototype network for few-shot 3D point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8883–8892, 2021.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3075–3084, 2019.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017.

Runyu Ding, Jihan Yang, Chuan-Sheng Xue, Wenwei Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-driven open-vocabulary 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7087–7096, 2023.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Yuxian Huang, Xidong Feng, Bo Li, Yuchuan Xiang, Haowei Wang, Ting Liu, and Bing Qin. Ensemble learning for heterogeneous large language models with deep parallel collaboration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 119838–119860, 2024.

Krishna Murthy Jatavallabhula, Alihusein Kuang, Zeng Lu, Qiao Gu, Jianning Ni, Sanat Kar, Maurice Fallon, John J. Leonard, and Liam Paull. ConceptFusion: Open-set multimodal 3D mapping with foundation models. In *Robotics: Science and Systems (RSS)*, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19729–19739, 2023.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2004.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 15682–15694, 2021.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. OpenScene: 3D scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3D semantic segmentation in the wild. In *European Conference on Computer Vision (ECCV)*, pp. 125–141, 2022.

Hao Tai, Qianyu He, Jing Zhang, Yiming Qian, Zhen Zhang, Xiaowei Hu, Xiang Li, Yaowei Wang, and Yong Liu. Open-vocabulary SAM3D: Towards training-free open-vocabulary 3D scene understanding. *arXiv preprint arXiv:2405.15580*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, pp. 23965–23998, 2022.

Yan Wu, Ming Xing, Yong Zhang, Yuan Xie, and Yanyun Qu. Fusion-then-distillation: Toward cross-modal positive distillation for domain adaptive 3D semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):9030–9045, 2025.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019.

Meng Yan, Jiaming Zhang, Yunkang Zhu, and Haifeng Wang. MaskClustering: View consensus based mask graph clustering for open-vocabulary 3D instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28274–28284, 2024.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19823–19832, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International Conference on Machine Learning (ICML)*, pp. 609–616, 2001.

Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Parameter is not all you need: Starting from non-parametric networks for 3D point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2105–2114, 2023.

Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3D point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8873–8882, 2021.

Yifan Zhao, Feng Zhao, Zhaoxiang Zhang, Shuang Wu, Yating Lin, and Yong Li. Learning hierarchical orthogonal prototypes for generalized few-shot 3D point cloud segmentation. *arXiv preprint arXiv:2603.19788*, 2026.

A Extended Robustness Diagnostics

A.1 Tie-Policy Robustness

On the same held-out 156 ScanNet200 scenes, 20,402,026 valid points have sparse evidence. Dense and sparse Top-1 disagree on 13,859,946 of them (67.934%). Table 7 reports deterministic tie policies. The frozen lowest-index policy is the strongest Top-1 variant, making the headline $+6.40$ gain conservative among the tested rules.

Table 7: Top-1 tie robustness. Full distribution is fixed at 34.87 HM.

Policy	Base	Novel	HM	Full − Top-1 [95% CI]
Frozen lowest-index	45.13	20.80	28.47	+6.40 $[+5.25, +7.65]$
Dense-favouring	43.60	19.06	26.53	+8.35 $[+7.15, +9.57]$
Sparse-favouring	23.98	17.18	20.02	+14.86 $[+13.41, +16.03]$

A.2 Fusion-Weight Sensitivity

We vary the sparse weight w_s while setting $w_d = 1 - w_s$. For the Top-1 control, the exact $w_s = 0.5$ tie is resolved in favor of dense to avoid class-index dependence. Full-distribution fusion is stable across the entire grid (Table 8). With one-hot sources, $w_s < 0.5$ makes dense win every disagreement and $w_s > 0.5$ makes sparse win every disagreement, so continuous source reweighting degenerates into a hard switch. Retained distributions preserve graded within-class evidence and therefore support genuinely graded weighting.

Table 8: Sparse-weight sensitivity; $w_d = 1 - w_s$. The $w_s = 0.5$ Top-1 tie is dense-favouring.

w_s	Top-1 HM	Full HM	ΔHM	Full Base	Full Novel
0.3	26.53	34.33	+7.80	46.07	27.36
0.4	26.53	34.74	+8.22	45.39	28.14
0.5	26.53	34.87	+8.35	44.43	28.70
0.6	20.02	34.64	+14.62	43.14	28.94
0.7	20.02	34.17	+14.15	41.60	28.99

A.3 Held-Out Calibration Diagnostics

Histogram calibrators (Zadrozny & Elkan, 2001) use 15 equal-width bins and are fitted only on the 156 fit scenes. Accuracy is unchanged by calibration; confidence statistics are evaluated on the disjoint held-out half. Calibration changes confidence reliability without changing source accuracy. The strong ECE reductions in Table 9 demonstrate that the fitted calibration machinery is effective, while the main ScanNet200 results show that calibrated probability fusion (34.51 HM) still does not exceed uniform full-distribution averaging (34.87 HM). This separates confidence miscalibration from irreversible loss of semantic alternatives.

Table 9: Source calibration on held-out ScanNet200. Accuracy, confidence, and ECE are percentages.

Source	State	Acc.	Mean conf.	ECE ↓	Brier ↓
Dense RegionPLC	raw	61.80	40.88	29.20	0.3117
Dense RegionPLC	calibrated	61.80	62.50	1.54	0.2146
Sparse SAM3	raw	38.02	81.31	43.29	0.4187
Sparse SAM3	calibrated	38.02	37.74	0.69	0.2190

A.4 GroundingDINO–SAM2.1 Retention Ladder

Table 10 provides the complete retention ladder on the five-scene source-replacement diagnostic with GroundingDINO–SAM2.1. The dense RegionPLC distribution is held fixed; only sparse Ground-

ingDINO semantic retention varies. The monotonic gain indicates that the retention phenomenon is not specific to SAM3.

Table 10: Five-scene source-replacement diagnostic. The dense RegionPLC distribution is fixed; only GroundingDINO semantic retention changes.

Retention	Base	Novel	HM	ΔHM vs. Top-1
Top-1	21.14	11.35	14.77	0.00
Top-2	22.62	11.91	15.61	+0.84
Top-5	24.20	12.54	16.52	+1.75
Top-10	25.39	12.94	17.14	+2.37
Top-20	26.22	13.80	18.08	+3.31
Full	27.56	14.21	18.75	+3.98

A.5 PoE Numerical-Floor Sensitivity

Geometric pooling requires a numerical floor because the sparse distribution contains exact zeros. Table 11 varies ϵ from 10^{-4} to 10^{-12} while keeping all source evidence and weights fixed. Harmonic Mean changes by only 0.16 points over the entire range and is effectively identical for $\epsilon \leq 10^{-8}$.

Table 11: Geometric/log pooling (Product of Experts) (Hinton, 2002) is stable across the numerical floor ϵ.

ϵ	Base	Novel	HM	ΔHM vs. 10^{-8}
10^{-4}	46.98	29.18	36.00	-0.15
10^{-6}	47.03	29.31	36.12	-0.04
10^{-8}	47.07	29.35	36.16	0.00
10^{-10}	47.07	29.35	36.16	-0.00
10^{-12}	47.07	29.35	36.16	-0.00

B Compute and Memory Accounting

The complete 312-scene probability cache contains 49,540,568 points. Its compressed NPZ footprint is 13.31 GiB; raw float16 dense distributions occupy 18.46 GiB and sparse COO evidence 2.93 GiB (315,063,992 entries). Dense storage costs 400 bytes per point and sparse COO 10 bytes per nonzero entry. The held-out 156 scenes contain 25,020,281 points and occupy 6.69 GiB compressed.

With chunk size 5,000, one 5000×200 float32 matrix is 3.81 MiB and three such matrices are approximately 11.44 MiB. Measured process peak RSS is 719.52 MiB. Full-distribution fusion of the held-out set takes 12.00 s on CPU (0.077 s/scene, 2,084,610 points/s); compressed NPZ loading takes 130.58 s and dominates wall-clock time. The largest cached scene contains 438,565 points; its unchunked $N \times 200$ float32 distribution would require 334.60 MiB.

C Qualitative Selection and Failure Analysis

Positive qualitative ScanNet200 scenes were selected only after quantitative evaluation using predeclared criteria: $\Delta\text{HM} > 0$, $\Delta\text{Novel} > 0$, and positive net point-level corrections. The selected scenes are `scene0356_01`, `scene0574_02`, and `scene0164_01`, with HM gains of $+15.49$, $+13.60$, and $+12.48$ respectively. The deliberately retained failure `scene0490_00` drops from 31.66 HM under matched Top-1 to 19.41 under Full, a -12.25 HM change with net correction count $-12{,}015$. These cases motivate adaptive retention rather than unrestricted tail preservation.