GSSR: BASE-PRESERVING GRAPH FILTERING FOR PROMPT-ROBUST 3D PART SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Meaning-preserving prompts for the same 3D part can produce different pointwise masks. We measure the graph-frequency structure of paraphrase-induced logit changes and find that normalized Dirichlet energy adds modest held-out predictive value for soft-mask disagreement across two fusion heads. We then test a selective placement of a standard graph filter: GSSR filters the final query-dependent logit residual while leaving the query-independent base unfiltered. For the same frozen fields and filter, residual-only and full-logit filtering act identically on prompt-pair logit differences; the placement question is whether preserving the base benefits segmentation. In a PTv3-based controlled model, GSSR lowers soft disagreement ($100e_{\mathrm{soft}}$) from 3.20 to 2.25 and raises worst-prompt mIoU from 44.60 to 48.80 relative to the unfiltered Direct control. Its primary 3DCoMPaT-Coarse mIoU is 51.17 versus 47.80 (five-seed means). Frozen and retrained placement controls show higher observed region and boundary scores than full-logit filtering; a second fusion head replicates the Direct/GSSR direction. These findings support selective filter placement in the tested model family.

1 INTRODUCTION

Given a 3D object and a prompt such as "Temple of eyeglasses," text-prompted part segmentation must determine what part is requested and where it lies. Language provides the semantic target, while geometry provides the spatial evidence. This suggests a simple contract: for a fixed object and target part, meaning-preserving rewordings should produce the same mask. In practice, however, equivalent prompts can induce spatially structured changes in point-wise logits and alter the predicted support. We refer to this failure as *prompt-equivalent spatial drift*.

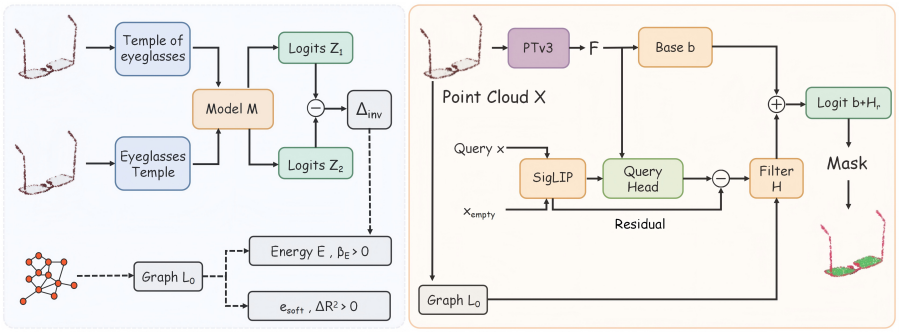
Recent systems transfer image-language priors, learn text-aligned point features, or introduce prompt-conditioned decoders Zhu et al. (2023); Zhou et al. (2023); Ma et al. (2025); Jin et al. (2026). These advances improve the representations being aligned and the mechanisms used to fuse them. However, they leave the spatial structure of the final language-induced update largely implicit. Consequently, a rewording may perturb where the model responds even when what the prompt refers to remains unchanged.

Filtering the complete logit can attenuate prompt-induced variation, but it also changes the query-independent base field. Sharp responses may reflect wording sensitivity or true part boundaries. We therefore ask whether filtering only the query-dependent update better preserves mask accuracy and boundary quality. We use fixed polynomial diffusion on the object graph, drawing on established graph-signal-processing tools (Shuman et al., 2013; Defferrard et al., 2016).

To motivate the intervention, we compare meaning-preserving prompt pairs whose desired masks coincide. We measure normalized graph Dirichlet energy of their pointwise logit differences and test whether it adds held-out predictive information about soft disagreement beyond nonspectral covariates. The association is modest across pooled-prompt and token–point heads; it is a diagnostic observation, not evidence that graph frequency causes the errors.

Graph-Spectral Semantic Residual (GSSR) operationalizes this comparison by filtering the empty-anchored query residual after text fusion while the shared base bypasses the operator (Figure 1). With identical frozen fields, residual-only and full-logit filtering produce the same prompt-pair logit differences; their outputs differ in the treatment of the base.

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.



(a) Prompt-equivalent diagnosis (b) GSSR framework

Figure 1: **Prompt-equivalent diagnosis and GSSR framework.** (a) For a fixed point cloud, meaning-preserving prompts should produce the same mask. We measure their logit difference on the geometry graph and test whether its normalized Dirichlet energy is associated with soft prediction disagreement. (b) GSSR forms an empty-anchored query residual and applies the fixed graph filter only after text fusion, while the query-independent base at inference bypasses the filter.

We first assess the diagnostic on held-out paraphrases, then compare GSSR with the unfiltered Direct control for prompt robustness and segmentation. Frozen same-residual and separately retrained controls address the residual-versus-full-logit placement question. A second fusion head tests the Direct/GSSR direction with the same PTv3 backbone, while published systems provide descriptive benchmark context.

Our contributions are threefold:

- We characterize prompt-equivalent spatial drift and observe that graph energy of paraphrase-induced logit differences adds modest held-out predictive value for soft-mask disagreement in two fusion heads.
- We evaluate a standard graph filter on the final query-dependent logit residual. For identical frozen fields, this and full-logit filtering have the same prompt-pair logit response, while only residual-only filtering leaves the base unchanged.
- Within one PTv3 backbone, we evaluate Direct/GSSR using paired training and a second fusion head, and test operator placement using frozen same-residual and separately retrained controls. External-method comparisons are reported only as descriptive system-level context.

2 Related Work

Language-guided 3D understanding. Language conditioned 3D understanding spans object grounding, scene segmentation, and fine-grained part prediction (Chen et al., 2020; Achlioptas et al., 2020; Chen et al., 2026). At the object level, PartGlot learns shape parts from referential language (Koo et al., 2022); PointCLIP V2 transfers image–language priors to point-cloud tasks (Zhu et al., 2023); and PartSLIP, PartSLIP++, and PartDistill lift or distill vision–language knowledge for 3D part segmentation (Liu et al., 2023; Zhou et al., 2023; Umam et al., 2024). Find3D learns scalable open-world part features, while CoSMo3D introduces canonical spatial modeling for promptable part semantics (Ma et al., 2025; Jin et al., 2026). At the scene level, OpenScene, PLA, and OpenMask3D associate open-vocabulary language features with dense points or 3D masks (Peng et al., 2023; Ding et al., 2023; Takmaz et al., 2023). These methods advance representation learning, transfer, and cross-modal fusion. GSSR is complementary: it controls the geometry-graph frequency content of the final prompt-dependent update without replacing the backbone or text encoder.

Prompt sensitivity and adaptation. Prompt and classifier adaptation improve vision–language transfer through input-conditioned contexts, test-time prompt optimization, or residual classifier updates (Zhou et al., 2022; Shu et al., 2022; Yu et al., 2023). PARC instead studies reliability under plausible visual and linguistic prompt variations (Schmalfuss et al., 2025). Our analysis fixes both

the object and semantic target, measures pointwise logit changes on the object geometry graph, and uses this output-level signal to motivate a post-fusion residual intervention.

Graph filtering and prediction refinement. Graph signal processing measures signal variation with a graph Laplacian and implements localized spectral responses with polynomial filters (Shuman et al., 2013; Defferrard et al., 2016). Related 3D methods diffuse surface features or propagate support labels (Sharp et al., 2022; Zhao et al., 2021), while Correct-and-Smooth and graph regularization refine predictions or training objectives (Huang et al., 2021; Yang et al., 2021). Frequency-aware vision–language methods operate on visual features, tokens, or two-dimensional region graphs (Gong et al., 2025; Ouyang et al., 2026; Huang et al., 2026). We test whether leaving the learned shared field unfiltered improves part-mask accuracy and boundary quality when both placements apply the same graph response to prompt-pair logit differences. We use the same graph and filter under these two placements and complement region mIoU with boundary-sensitive evaluation (Cheng et al., 2021; Tang et al., 2022); exact diagnostic graph construction appears in Appendix A.1, and the boundary-metric graph is defined in Appendix B.2.

3 METHODOLOGY

We fix a point cloud and compare prompts designated as meaning-preserving references to the same part. Their desired masks coincide, so differences between their predictions expose prompt sensitivity. We first measure the spatial structure of this variation, then identify the final query-dependent logit update, and finally specify where an established geometry-graph filter acts. Our method claim concerns that placement, not a new spectral operator.

3.1 PROMPT-EQUIVALENT OUTPUT VARIATION

Let

$$\mathcal{X} = \{(\mathbf{p}_i, \mathbf{a}_i)\}_{i=1}^{N}$$

be a point cloud and let x be a text prompt with binary target $y^x \in \{0,1\}^N$. A model M produces pointwise logits $\ell_M(\mathcal{X}, x) \in \mathbb{R}^N$ and probabilities $p_M(\mathcal{X}, x) = \mathrm{sigmoid}(\ell_M(\mathcal{X}, x))$. For a meaning-preserving pair (x, x^+) that names the same valid part, we define

$$\begin{aligned} \Delta_{\mathrm{inv}} &= \ell_M(\mathcal{X}, x) - \ell_M(\mathcal{X}, x^+), \\ e_{\mathrm{soft}} &= \frac{1}{N} \sum_{i=1}^{N} \left| p_{M,i}(\mathcal{X}, x) - p_{M,i}(\mathcal{X}, x^+) \right|. \end{aligned} \tag{1}$$

The logit difference records where the model response changes, while e_{soft} records the mean change in output probability. These output-level quantities apply regardless of where text enters the network.

We build a symmetric k_g-nearest-neighbor graph from normalized point coordinates and denote its normalized Laplacian by L_0. For a graph signal Z, normalized Dirichlet energy is

$$\mathcal{E}_{L_0}(Z) = \frac{\mathrm{tr}(Z^\top L_0 Z)}{\|Z\|_F^2 + \epsilon}. \tag{2}$$

We use $\epsilon = 10^{-12}$ and treat signals with $\|Z\|_F / \sqrt{N} < 10^{-6}$ as undefined. The statistic measures edge-wise variation relative to signal magnitude. High energy is not an error label by itself: true part boundaries and thin structures can also vary sharply. Prompt-equivalent pairs make the diagnostic specific, because their desired mask difference is zero.

We test on held-out objects whether this graph energy adds predictive information about soft disagreement beyond prespecified nonspectral covariates. Section 4.3 reports the grouped out-of-fold comparison, while Appendix A.1 gives the regression, aggregation, and interval procedure. An association here motivates a placement test but does not establish a causal source of error.

3.2 OPERATIONAL QUERY RESIDUAL

A text-independent backbone produces $F = B(\mathcal{X}) \in \mathbb{R}^{N \times d}$. We separate its final prediction into a prompt-independent base field $b_\phi(F)$ and a raw query field $a_\theta(F, x)$. A fixed empty prompt $x_\emptyset$

defines the operational language residual

$$r_\theta(F, x) = a_\theta(F, x) - a_\theta(F, x_\emptyset), \qquad r_\theta(F, x_\emptyset) = 0. \tag{3}$$

For controlled comparisons, we define Direct as the matched identity control, with logits

$$\ell_{\mathrm{Direct}}(F, x) = b_\phi(F) + r_\theta(F, x). \tag{4}$$

Direct replaces the GSSR operator by I_N while retaining the same model family. The empty string is an operational reference, not a verified absent-part prompt or an identified semantic null class. It is shared by Direct and GSSR and cancels from any prompt-pair residual difference:

$$r_\theta(F, x) - r_\theta(F, x') = a_\theta(F, x) - a_\theta(F, x').$$

The anchor therefore does not itself attenuate prompt-to-prompt logit changes. Both controlled variants hold this decomposition fixed; the independent design choice is whether to filter r_θ before adding the query-independent base. The frozen controls in Section 4.5 compare that placement directly.

3.3 Residual-only graph filtering

We use "graph filtering" for the general operation and "graph diffusion" for the particular polynomial response below. GSSR uses the diagnosis geometry graph, with $L = L_0$, and defines

$$H_{\eta,m} = (I_N - \eta L)^m, \qquad 0 < \eta \le \tfrac{1}{2}, \quad m \in \mathbb{N}_+. \tag{5}$$

The reported model uses $(k_g, \eta, m) = (16, 0.25, 4)$, fixed on the design split. We write $H = H_{\eta,m}$ when unambiguous. Its final logit is

$$\ell_{\mathrm{GSSR}}(F, x) = b_\phi(F) + H_{\eta,m} r_\theta(F, x). \tag{6}$$

All text interaction occurs before $H_{\eta,m}$, and no query-conditioned operation follows it. One application of $I_N - \eta L$ mixes a residual response with neighboring graph responses; repeating it m times suppresses components that vary rapidly across edges. The learned base bypasses these iterations. The filter adds no trainable parameters.

For a Laplacian eigenvalue $\lambda \in [0, 2]$, the frequency response is $h(\lambda) = (1 - \eta\lambda)^m$, which is nonnegative and monotonically decreasing. For $\lambda_c \in [0, 2]$, if $\Pi_{\ge\lambda_c}$ projects onto eigenvectors with eigenvalue at least λ_c, then for any residual graph signal r

$$\left\|\Pi_{\ge\lambda_c} H_{\eta,m} r\right\|_2 \le (1 - \eta\lambda_c)^m \left\|\Pi_{\ge\lambda_c} r\right\|_2. \tag{7}$$

Moreover, any two prompts evaluated on the same point cloud satisfy

$$\ell_{\mathrm{GSSR}}(F, x) - \ell_{\mathrm{GSSR}}(F, x') = H_{\eta,m}\left[a_\theta(F, x) - a_\theta(F, x')\right]. \tag{8}$$

For brevity, write $b = b_\phi(F)$, $r(F, x) = r_\theta(F, x)$, and $a(F, x) = a_\theta(F, x)$. For the same frozen fields and H, full-logit filtering gives $\ell_{\mathrm{Full}}(F, x) = H[b + r(F, x)]$. Its difference between two prompts is also $H[r(F, x) - r(F, x')] = H[a(F, x) - a(F, x')]$. The logits for one prompt differ by

$$\ell_{\mathrm{GSSR}}(F, x) - \ell_{\mathrm{Full}}(F, x) = (I_N - H)b.$$

Thus Eq. equation 7 is a standard pre-sigmoid graph-filter bound, and Eq. equation 8 does not distinguish residual-only from full-logit placement. The testable hypothesis is whether preserving the learned query-independent base improves masks or boundaries under identical prompt-pair logit filtering. Neither equation guarantees lower probability disagreement, correct thresholded masks, or boundary fidelity. The exact equality need not hold across independently retrained models, whose learned b and r can adapt differently.

3.4 Controlled model instantiation

For each query, the text-independent backbone first computes point features F and a base logit $b_\phi(F)$. The query head is evaluated on the requested text and on the empty reference; subtracting its two outputs forms $r_\theta(F, x)$. GSSR filters this residual on the geometry graph and adds back $b_\phi(F)$ before the sigmoid. Direct follows the same steps with $H = I_N$.

The primary controlled model uses PTv3 (Wu et al., 2024), a frozen SigLIP text encoder (Zhai et al., 2023), and pooled-prompt modulation. Direct and GSSR share the backbone and fusion-head architectures, task loss, optimization schedule, initialization protocol, and trainable parameter count. They are trained independently under paired seed identities and differ by I_N versus H in the final residual path. The token–point study changes the fusion head but retains PTv3; it is a cross-head test, not a cross-backbone test.

All variants optimize binary cross-entropy plus Dice loss on valid part queries. GSSR uses neither a spectral auxiliary loss nor boundary supervision, and absent prompts are not training examples. Head and task-loss definitions appear in Appendix A.2; the placement-control operators are defined in Appendix B.3.

4 Experiments

4.1 Experimental setup and comparison roles

Controlled models are trained on 3DCoMPaT200 (Ahmed et al., 2025) and evaluated on the source-domain 3DCoMPaT-Coarse and 3DCoMPaT-Fine benchmarks (Ahmed et al., 2025; Jin et al., 2026) and the cross-dataset ShapeNet-Part (Yi et al., 2016) and PartNet-E (Liu et al., 2023; Mo et al., 2019) transfer sets. A category-stratified manifest separates training, design, and held-out confirmation data. The confirmation split is used for the counterfactual diagnosis and the matched prompt-equivalence evaluation; none of these outcomes is used for parameter fitting, checkpoint selection, or graph/filter selection. The remaining segmentation rows use the named benchmark test sets.

We assign each comparison a distinct role. PointCLIP V2, PartSLIP++, OpenMask3D, Find3D, and CoSMo3D are published external baselines (Zhu et al., 2023; Zhou et al., 2023; Takmaz et al., 2023; Ma et al., 2025; Jin et al., 2026). Direct is instead the matched identity control within our model family. Direct and GSSR use the same backbone and fusion-head architectures, data, loss, optimization schedule, and paired seed identities, but they are optimized independently and differ in using I_N or $H_{\eta,m}$ on the final residual. Frozen and retrained alternatives are internal placement controls.

The GSSR row in the external comparison is the same five-seed result used in the matched Direct/GSSR study. The two tables therefore reuse one GSSR record: Table 1 compares that record descriptively with published systems, whereas Table 3 compares it with Direct under the matched model-family protocol.

Published baselines use the aggregate results and method-specific protocols reported by their original sources, for which run-level outcomes are unavailable. Our GSSR entry is a five-seed mean. The controlled pooled-prompt Direct/GSSR comparison uses five paired training seed identities. Frozen interventions reuse the same five Direct checkpoints; independently retrained placement controls and the token–point replication each use separate three-seed protocols. Unless stated otherwise, scores are object-macro percentages reported as mean $\pm$ seed standard deviation. For controlled studies with retained seed-level outcomes, paired differences use the training seed as the inferential unit and descriptive 95% Student-t intervals are reported where available. External-method comparisons use reported aggregate scores and receive no paired inference.

The primary controlled endpoint is 3DCoMPaT-Coarse mIoU under reference orientation and object-aware prompts. We additionally report tolerant boundary F1 (BF1), strict boundary IoU (BIoU_{16}), thin-part mIoU, and transfer performance. Boundary eligibility is fixed from ground truth rather than selected by model predictions. BF1 is a secondary outcome; transfer, macro-average, BIoU, and thin-part results are unadjusted descriptive outcomes. Appendix B.1 gives the split, aggregation, and uncertainty definitions; Appendix B.2 gives boundary-metric eligibility and construction.

4.2 Comparison with published methods

Table 1 positions GSSR against published systems under reference orientation with object-aware prompts. The complete prompt-by-orientation grid is reported in Appendix D.

Table 1: Comparison under the canonical-orientation and object-aware-prompt setting, "{part} of a {object}". We report object-macro mIoU (%). Published baseline values are taken from Table 1 of CoSMo3D (Jin et al., 2026), whereas the GSSR row reports our five-seed mean under the same evaluation setting. Because the published methods retain their own training pipelines and run-level outcomes are unavailable, cross-method differences are descriptive. "–" denotes an unreported result.

Method	3DCoMPaT-C	3DCoMPaT-F	ShapeNet-Part	PartNet-E
OpenMask3D	–	–	8.94	12.54
PointCLIP V2	14.09	7.25	16.91	11.28
PartSLIP++	6.12	3.79	–	–
Find3D	31.72	12.69	28.39	16.86
CoSMo3D	47.51	24.48	36.16	17.59
GSSR	**51.17**	**28.66**	**38.71**	**19.52**

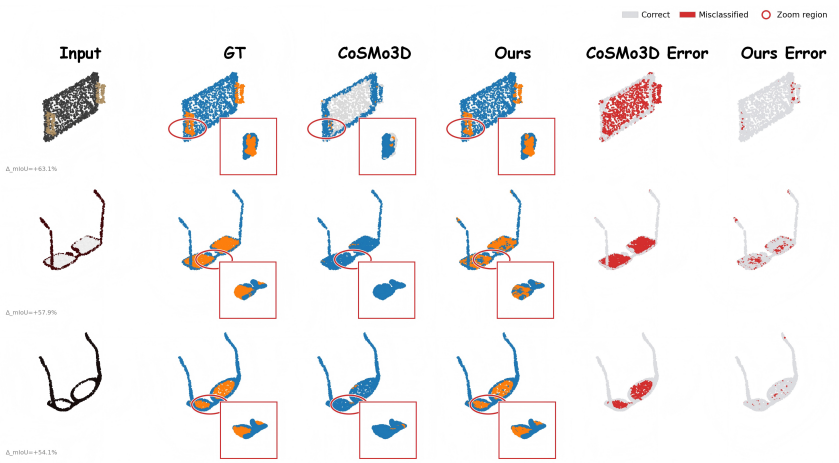


Figure 2: **Qualitative part-segmentation results.** Columns show the input point cloud, ground truth, CoSMo3D and GSSR predictions, and their corresponding error maps; "Ours" in the image denotes GSSR. Red points mark misclassified points, and the outlined regions are enlarged in the insets. The image labels give the per-example GSSR-minus-CoSMo3D mIoU differences in percentage points. These complete systems retain their respective training and inference pipelines, so the examples are descriptive rather than isolated estimates of the residual-filtering effect.

GSSR has the largest reported mean among the available method entries in all four displayed benchmarks and all 16 benchmark–prompt–orientation settings in the complete grid. Aggregate scores, however, do not show where the predictions differ. Figure 2 therefore compares representative point-level outputs from CoSMo3D and GSSR, including enlarged part regions and the corresponding error maps.

Together, Table 1 and Figure 2 position GSSR as a complete system but do not identify the source of its improvement. We therefore reuse the same GSSR benchmark record below and change the comparator: Direct supplies the controlled identity-placement reference, while frozen and retrained controls test the proposed operator placement.

4.3 Graph energy tracks prompt-equivalent disagreement

The external comparison does not explain which component of GSSR matters. Table 2 first evaluates the output-level diagnosis on the held-out confirmation split. The pooled-prompt analysis uses the five Direct checkpoints from the matched study, and the token–point analysis uses three independently trained Direct checkpoints.

Graph energy has positive coefficient and held-out increment estimates in both fusion heads. The increments are modest but consistent: spatial variation in prompt-equivalent logit differences carries

Table 2: Held-out counterfactual diagnosis. β_E is the coefficient of the fold-standardized energy predictor, and ΔR^2 is its out-of-fold increment over the prespecified nonspectral covariates. Intervals are descriptive 95% intervals; folds are grouped by object, and results are conditional on rows with defined energy.

Head	β_E [95% CI]	ΔR^2 [95% CI]
Pooled prompt (5 seeds)	$0.29\,[0.17, 0.40]$	$0.06\,[0.03, 0.09]$
Token–point (3 seeds)	$0.25\,[0.12, 0.37]$	$0.05\,[0.02, 0.08]$

Table 3: Matched Direct/GSSR robustness and segmentation comparison. The top panel evaluates held-out object–part paraphrase groups using five paired seeds; scores are averaged first within each paraphrase group and then over objects. The remaining panels report reference-orientation, object-aware-prompt benchmark scores. Δ denotes GSSR minus Direct. Bracketed values are paired descriptive 95% confidence intervals over five training-seed pairs for the prompt-robustness measures and primary Coarse mIoU. Unbracketed differences are differences of reported means. Seed SD is shown where available.

Measure	Direct ($H = I_N$)	GSSR ($H = H_{\eta,m}$)	Δ (GSSR$-$Direct)
Held-out prompt-equivalent robustness			
Soft disagreement ($100\, e_{\mathrm{soft}}$) $\downarrow$	3.20	**2.25**	$\mathbf{-0.95}\,[\mathbf{-1.25}, \mathbf{-0.65}]$
Point-flip rate (%) $\downarrow$	2.85	**1.95**	$\mathbf{-0.90}\,[\mathbf{-1.20}, \mathbf{-0.60}]$
Normalized Dirichlet energy $\downarrow$	0.310	**0.205**	$\mathbf{-0.105}\,[\mathbf{-0.135}, \mathbf{-0.075}]$
Worst-prompt mIoU (%) $\uparrow$	44.60	**48.80**	$\mathbf{+4.20}\,[\mathbf{+3.10}, \mathbf{+5.30}]$
Overall object-macro mIoU			
3DCoMPaT-Coarse (primary)	47.8 ± 0.6	$\mathbf{51.17 \pm 0.6}$	$\mathbf{+3.37}\,[\mathbf{+2.57}, \mathbf{+4.17}]$
3DCoMPaT-Fine	24.70	**28.66**	**+3.96**
ShapeNet-Part	36.30	**38.71**	**+2.41**
PartNet-E	17.70	**19.52**	**+1.82**
Transfer macro	27.00	**29.12**	**+2.12**
Four-dataset macro	31.63	**34.52**	**+2.89**
3DCoMPaT-Coarse boundary and thin-part scores			
Tolerant BF1 ($t = 1$)	41.3 ± 0.7	$\mathbf{43.8 \pm 0.6}$	$\mathbf{+2.5}$
BIoU_{16}	7.60	**16.11**	**+8.51**
Thin-part mIoU	21.5 ± 0.7	$\mathbf{23.0 \pm 0.6}$	$\mathbf{+1.5}$

information about undesirable prediction disagreement beyond the prespecified covariates. This association motivates the intervention target but does not establish graph frequency as a causal source of segmentation error. Additional energy and absent-query results are reported in the supplementary material.

4.4 Direct Robustness and Matched Segmentation Effects

Table 2 establishes that graph-frequency energy carries held-out information about prompt-equivalent disagreement within Direct models, but it does not show that GSSR reduces this failure mode. We therefore compare Direct and GSSR on the same held-out paraphrase groups. Prompt-pair measurements are averaged within each object–part group before object-level macro-aggregation. The top panel of Table 3 reports soft disagreement as the primary endpoint, point-flip rate, normalized graph energy, and worst-prompt mIoU; the last metric guards against trivially consistent but incorrect predictions. Direct and GSSR share the learned-component architecture and training protocol and differ only in whether the final query-dependent logit residual passes through I_N or $H_{\eta,m}$. Paired seed-level outcomes are available for the prompt-robustness measures and the primary 3DCoMPaT-Coarse mIoU endpoint, for which Table 3 reports paired descriptive 95% confidence intervals over five training-seed pairs. The remaining benchmark mIoU and cross-dataset macro results are reported as aggregate means without paired intervals.

The top panel directly evaluates the failure mode that motivates GSSR. Relative to Direct, GSSR reduces $100\, e_{\mathrm{soft}}$ from 3.20 to 2.25, point-flip rate from 2.85% to 1.95%, and normalized

Table 4: Complementary placement controls on 3DCoMPaT-Coarse under reference orientation and object-aware prompts. Panel A replaces the operator on identical residuals from five frozen Direct checkpoints and uses the common validity set. Panel B independently retrains each variant with three paired seeds. Values are mean ± seed SD and are interpreted within each panel.

A. Frozen same-residual (5 seeds)			*B. Independently retrained (3 seeds)*		
Output	mIoU	BF1	Variant	mIoU	BF1
Direct ($b + r$)	47.2 ± 0.4	41.4 ± 0.6	Direct	47.6 ± 0.5	41.6 ± 0.7
Residual low-pass	$\mathbf{47.8 \pm 0.4}$	$\mathbf{43.1 \pm 0.5}$	**Residual low-pass**	$\mathbf{50.2 \pm 0.4}$	$\mathbf{43.9 \pm 0.6}$
Norm-matched scalar	47.3 ± 0.4	41.8 ± 0.6	Residual high-pass+ker L	47.1 ± 0.5	40.9 ± 0.7
Residual high-pass+ker L	47.1 ± 0.4	41.2 ± 0.6	Full-logit low-pass	47.9 ± 0.5	42.0 ± 0.7
Shuffled-graph low-pass	47.0 ± 0.4	41.0 ± 0.6	Laplacian regularization	48.0 ± 0.5	42.1 ± 0.7
Full-logit low-pass	47.1 ± 0.4	41.5 ± 0.6			

Dirichlet energy from 0.310 to 0.205. All three paired descriptive intervals exclude zero. At the same time, worst-prompt mIoU increases from 44.60 to 48.80, with a paired difference of $+4.20\,[+3.10, +5.30]$. Thus, the reduction in disagreement is accompanied by improved worst-case prompt accuracy rather than a trivially consistent empty-mask solution. Additional diagnostic energy and absent-query false-positive-rate outcomes appear in Appendix C.1; frozen same-residual controls are summarized in Table 4. Having verified the intended prompt-robustness effect directly, we next examine whether it co-occurs with improvements in conventional segmentation and boundary quality.

On the primary 3DCoMPaT-Coarse endpoint, GSSR achieves 51.17 ± 0.60 mIoU, compared with 47.80 ± 0.60 for Direct, with a paired mean improvement of 3.37 points (descriptive 95% CI: $[2.57, 4.17]$). The four-dataset macro mean is 2.89 points higher, and the mean is also higher on both transfer sets. GSSR has higher observed mean BF1, BIoU_{16}, and thin-part mIoU by 2.5, 8.51, and 1.5 points, respectively. These cross-dataset macro, transfer, boundary, and thin-part contrasts are reported as descriptive mean differences without paired intervals. Direct and GSSR each have 46.7M trainable parameters; GSSR increases batch-1 latency from 69 to 73 ms. Full timing details appear in the supplementary material.

4.5 Frozen residuals provide a conditional operator contrast

Matched training compares complete learned models, so the observed difference can reflect both operator placement and adaptation during optimization. We next evaluate an inference-time contrast by replacing the operator on identical residuals from frozen Direct checkpoints. This comparison is conditional on the retained operator-valid set. The controls match update magnitude, reverse the frequency response, shuffle graph geometry, or diffuse the complete logit. Exact operators and exclusion rules are given in Appendix B.3.

Panel A of Table 4 reports the common operator-valid intersection. The Direct values are recomputed on this conditional set and therefore should not be compared numerically with Table 3.

Residual low-pass filtering has the largest observed mean mIoU and BF1, with increases of 0.6 and 1.7 points over Direct. Full-logit low-pass filtering gives 47.1 mIoU and 41.5 BF1, versus 47.8 and 43.1 for residual-only filtering. For the same frozen fields and H, both filtered variants have identical prompt-pair logit differences; their output-logit difference is $(I_N - H)b$. The observed gap is therefore consistent with a base-routing effect on this conditional subset, but the result does not identify which components of b contain useful boundaries. The norm-matched scalar has a smaller BF1 increase, whereas the high-pass and shuffled-graph rows are below Direct in BF1. Excluded tuple counts were not retained, limiting a coverage audit.

4.6 Retrained controls test placement after adaptation

Frozen replacements exclude adaptation during training. We therefore independently retrain Direct, residual low-pass, residual high-pass, full-logit diffusion, and Laplacian-regularized alternatives under a separate three-seed protocol. Panel B of Table 4 compares only models within that protocol.

Residual-only filtering again obtains the largest observed means, with mIoU and BF1 increases over Direct of 2.6 and 2.3 points. Full-logit diffusion and Laplacian regularization have mIoU increases of 0.3 and 0.4 points, compared with 2.6 points for residual-only filtering. The two panels estimate different effects: an immediate intervention on identical fields and a separately learned architecture comparison. In the latter, the models' b and r may adapt differently, so the frozen-model identity does not equate their prompt-pair logit differences. Neither panel compares against a full-logit filter selected with a separate equal-budget search; the three-seed placement gaps are descriptive.

4.7 Cross-head replication preserves the matched direction

The controlled panels share one PTv3 model family. A separate three-seed token–point replication therefore tests the Direct/GSSR comparison with a different fusion head. Mean four-dataset macro mIoU is 31.2 ± 0.5 for Direct and 32.1 ± 0.4 for GSSR, an estimated paired increase of $0.9\,[0.2, 1.6]$ points; mean BF1 is also higher, from 40.9 ± 0.9 to 43.0 ± 0.8. These three-seed estimates are descriptive. Full results are reported in the supplementary material.

5 Limitations

Graph variation can reflect true boundaries, noise, nonuniform sampling, or incorrect edges. $H_{\eta,m}$ is a soft diffusion-based graph filter, not an orthogonal projector. Its attenuation guarantee concerns pre-sigmoid prompt-dependent logit components and does not ensure a correct binary mask. For identical frozen fields, full-logit and residual-only filtering give the same prompt-pair logit differences; the proposed advantage concerns the unfiltered query-independent base. This learned base is not guaranteed to contain correct boundary evidence, and we do not measure whether the difference field $(I_N - H)b$ is localized at true boundaries.

The controlled experiments use one PTv3 backbone, one operational anchor, and one design-selected graph/filter configuration. The full-logit placement control shares that configuration rather than receiving a separate equal-budget filter search. The paired interval for the primary Coarse mIoU contrast is based on only five training-seed pairs, limiting the precision of uncertainty estimation. Paired intervals are not reported for the remaining pooled-prompt benchmark mIoU, cross-dataset macro, boundary, and thin-part contrasts. The frozen comparison is conditional on a common operator-valid subset whose excluded counts were not retained. Published baselines use their own pipelines and are descriptive complete-system comparisons. Additional backbone, anchor, graph-sensitivity, and boundary-localized mechanism studies are needed for broader claims.

6 Conclusion

We characterized prompt-equivalent spatial drift in text-prompted 3D part segmentation. Graph energy of paraphrase-induced logit changes adds modest held-out predictive value for soft-mask disagreement in two PTv3 fusion heads. GSSR applies a standard graph filter to the final query-dependent logit residual and leaves the learned query-independent base unchanged. For fixed model fields, residual-only and full-logit filtering give identical prompt-pair logit differences; their distinction is whether the base is filtered.

In matched PTv3 training, GSSR reduces $100e_{\mathrm{soft}}$ from 3.20 to 2.25 and raises worst-prompt mIoU from 44.60 to 48.80 relative to Direct. The primary Coarse mIoU means are 51.17 and 47.80, respectively, with a paired mean improvement of 3.37 points (descriptive 95% CI: $[2.57, 4.17]$). Frozen and separately retrained placement controls show higher observed mIoU and BF1 means for residual-only than for full-logit filtering, while a second fusion head replicates the Direct/GSSR direction. These results support a base-preserving design choice within the evaluated model family. Other backbones, anchors, filter settings, and independently tuned full-logit controls remain open tests of its scope.

AI use statement

We used ChatGPT solely to improve the grammar, clarity, and readability of the manuscript. The authors reviewed and edited all AI-assisted changes and verified that the technical content, results,

citations, and conclusions were preserved. The authors take responsibility for the final content of this paper.

Reproducibility statement

Section 3 and Appendix A describe the query-residual construction, graph filter, fusion heads, and training loss. Section 4.1 and Appendix B document the datasets, the roles of the training, design, and held-out confirmation splits, the paired-seed comparison protocol, evaluation metrics, and statistical reporting. Additional controlled results are provided in Appendix C.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *European Conference on Computer Vision*, pp. 422–440, 2020.

Mahmoud Ahmed, Xiang Li, Arpit Prajapati, and Mohamed Elhoseiny. 3DCoMPaT200: Language-grounded compositional understanding of parts and materials of 3D shapes. *arXiv preprint arXiv:2501.06785*, 2025.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, pp. 202–221, 2020.

Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, and Liujuan Cao. 3D-DRES: Detailed 3D referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2966–2974, 2026. doi: 10.1609/aaai.v40i4.37288.

Bowen Cheng, Ross B. Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3837–3845, 2016.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing 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*, pp. 7010–7019, 2023.

Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. FE-CLIP: Frequency enhanced CLIP model for zero-shot anomaly detection and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21220–21230, 2025.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations*, 2021.

Qiming Huang, Hao Ai, and Jianbo Jiao. Structure-aware feature rectification with region adjacency graphs for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4118–4127, 2026.

Li Jin, Weikai Chen, Yujie Wang, Yingda Yin, Zeyu Hu, Runze Zhang, Keyang Luo, Shengju Qian, Xin Wang, and Xueying Qin. CoSMo3D: Open-world promptable 3D semantic segmentation through LLM-guided canonical spatial modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14325–14334, 2026.

Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J. Guibas, and Minhyuk Sung. PartGlot: Learning shape part segmentation from language reference games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16505–16514, 2022.

Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. PartSLIP: Low-shot part segmentation for 3D point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21736–21746, 2023.

Ziqi Ma, Yisong Yue, and Georgia Gkioxari. Find any part in 3D. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7818–7827, 2025.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 909–918, 2019.

Shuyi Ouyang, Gongfan Fang, Xinyin Ma, Yen-Wei Chen, Lanfen Lin, and Xinchao Wang. Language-guided frequency modulation for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39175–39182, 2026.

Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

Jenny Schmalfuss, Nadine Chang, Vibashan VS, Maying Shen, Andrés Bruhn, and Jose M. Alvarez. PARC: A quantitative framework uncovering the symmetries within vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25081–25091, 2025.

Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. DiffusionNet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics*, 41(3):27:1–27:16, 2022. doi: 10.1145/3507905.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14274–14289, 2022.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013. doi: 10.1109/MSP.2012.2235192.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68367–68390, 2023.

Liyao Tang, Yibing Zhan, Zhe Chen, Baosheng Yu, and Dacheng Tao. Contrastive boundary learning for point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8489–8499, 2022.

Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. PartDistill: 3D shape part segmentation by vision-language model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3470–3479, 2024.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4840–4851, 2024.

Han Yang, Kaili Ma, and James Cheng. Rethinking graph regularization for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4573–4581, 2021. doi: 10.1609/aaai.v35i5.16586.

Li Yi, Vladimir G. Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas J. Guibas. A scalable active framework for region annotation in 3D shape collections. *ACM Transactions on Graphics*, 35(6):210:1–210:12, 2016. doi: 10.1145/2980179.2980238.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 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*, pp. 8873–8882, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022.

Yuchen Zhou, Jiayuan Gu, Xuanlin Li, Minghua Liu, Yunhao Fang, and Hao Su. PartSLIP++: Enhancing low-shot 3d part segmentation via multi-view instance segmentation and maximum likelihood estimation. *arXiv preprint arXiv:2312.03015*, 2023.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2639–2650, 2023.

A Additional method details

A.1 Counterfactual analysis implementation

Graph construction. For object centroid $\overline{\mathbf{p}}$, write $\widetilde{\mathbf{p}}_i = (\mathbf{p}_i - \overline{\mathbf{p}})/s_{\mathcal{X}}$, where $s_{\mathcal{X}} > 0$ is the fixed rotation-invariant scalar used by the isotropic coordinate-normalization pipeline. Each k_g-NN neighborhood, with $k_g = 16$, excludes its query point and resolves distance ties by ascending point index; all evaluated point clouds satisfy $N \geq 17$. We set

$$\sigma_0 = \frac{1}{N k_g} \sum_i \sum_{j \in \mathcal{N}_{k_g}(i)} \|\widetilde{\mathbf{p}}_i - \widetilde{\mathbf{p}}_j\|_2 + 10^{-8}$$

and

$$(A_0^{\mathrm{dir}})_{ij} = \begin{cases} \exp\Big(-\|\widetilde{\mathbf{p}}_i - \widetilde{\mathbf{p}}_j\|_2^2 / (2\sigma_0^2)\Big), & j \in \mathcal{N}_{k_g}(i), \\ 0, & \text{otherwise.} \end{cases}$$

The symmetric affinity is $(A_0)_{ij} = (A_0)_{ji} = \max\{(A_0^{\mathrm{dir}})_{ij}, (A_0^{\mathrm{dir}})_{ji}\}$. We assume no isolated vertices, set $(D_0)_{ii} = \sum_j (A_0)_{ij}$, and use

$$L_0 = I_N - D_0^{-1/2} A_0 D_0^{-1/2}, \qquad \mathcal{E}_{L_0}(Z) = \frac{\mathrm{tr}(Z^\top L_0 Z)}{\|Z\|_F^2 + \epsilon}.$$

We use $\epsilon = 10^{-12}$. Signals satisfying $\|Z\|_F / \sqrt{N} < 10^{-6}$ are treated as undefined rather than assigned zero.

Absent-query metric. For an absent prompt x^-, point false-positive rate at $\tau = 0.5$ is

$$\mathrm{FPR}(\mathcal{X}, x^-) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[p_{M,i}(\mathcal{X}, x^-) \geq \tau].$$

Valid and absent queries are first averaged within object and then macro-averaged over objects. Absent queries are excluded from valid-query mIoU and boundary metrics.

Regression covariates and aggregation. For seed s and cross-validation fold f, we fit the following ordinary least-squares model on the training rows of that fold:

$$e_{\mathrm{soft},ops} = \alpha_{sf} + \beta_{E,sf}\widetilde{E}_{opsf} + \boldsymbol{\gamma}_{sf}^{\top}\widetilde{\mathbf{q}}_{opsf} + \delta_{sf,\mathrm{cat}(o)} + \varepsilon_{opsf}. \tag{9}$$

Here o indexes objects, p indexes paraphrase pairs, $E_{ops} = \mathcal{E}_{L_{0,o}}(\Delta_{\mathrm{inv},ops})$, and $\delta_{sf,\mathrm{cat}(o)}$ is a reference-coded category fixed effect with one category omitted. The analysis-protocol non-energy vector $\mathbf{q}$ contains $\|\Delta_{\mathrm{inv}}\|_2/\sqrt{N}$, mean and maximum absolute logit difference, mean threshold margin, confidence, entropy, valid-part fraction, boundary-to-area ratio, fixed-graph edge-label agreement, and the mean IoU of the two paraphrase predictions against ground truth. Energy and every continuous covariate are standardized using only the corresponding training fold. For continuous covariate dimension j,

$$\widetilde{q}_{j,opsf} = \frac{q_{j,ops} - \mu_{q_j,sf}}{\sigma_{q_j,sf}}.$$

A zero-variance covariate dimension is omitted from both the full and reduced models for that fold.

Five-fold cross-validation is grouped by object and stratified by category. The full and reduced models share the same folds, every object–pair row for an object remains in one fold, and each retained row receives equal weight. Rows with undefined energy are removed from both models, so each seed uses one common complete-case set. For training-fold mean $\mu_{E,sf}$ and standard deviation $\sigma_{E,sf} > 0$,

$$\widetilde{E}_{opsf} = \frac{E_{ops} - \mu_{E,sf}}{\sigma_{E,sf}}.$$

The seed-level coefficient is the arithmetic mean of the five fold-specific $\beta_{E,sf}$ estimates. Concatenating held-out predictions across folds gives

$$R^2_{\mathrm{OOF},s} = 1 - \frac{\sum_{o,p}(e_{\mathrm{soft},ops} - \widehat{e}^{\,\mathrm{OOF}}_{\mathrm{soft},ops})^2}{\sum_{o,p}(e_{\mathrm{soft},ops} - \overline{e}_{\mathrm{soft},s})^2},$$

where $\overline{e}_{\mathrm{soft},s}$ is the mean over all retained object–pair rows for seed s. Coefficients and out-of-fold R^2 increments are finally averaged arithmetically over seeds.

The same confirmation objects are evaluated by every model seed. Intervals therefore use a crossed seed/object bootstrap: each replicate resamples seed IDs and a common set of object IDs, retains all prompt-pair rows for each selected seed–object combination, and reruns fold standardization, the full and reduced regressions, and grouped cross-validation. The same hierarchy is used for absent FPR. Because only five pooled-prompt and three token–point seeds are available, these intervals remain descriptive.

A.2 Fusion-head implementation

Text is encoded by the frozen encoder and tokenizer. The empty reference is the tokenizer output for the empty string, including standard special tokens. We use $d = d_t = 768$, hidden width $d_h = 256$, and maximum text length 64.

Base and pooled-prompt heads. The base field is

$$b_{\phi,i}(F) = \mathbf{w}_b^{\top}\,\mathrm{GELU}(W_b\mathbf{f}_i + \mathbf{c}_b) + c_{\mathrm{out}},$$

where $W_b \in \mathbb{R}^{d_h\times d}$, $\mathbf{w}_b, \mathbf{c}_b \in \mathbb{R}^{d_h}$, and $c_{\mathrm{out}} \in \mathbb{R}$. The pooled-prompt head uses the normalized pooled text output $\mathbf{x} \in \mathbb{R}^{d_t}$, with $W_f \in \mathbb{R}^{d_h\times d}$, $W_x \in \mathbb{R}^{d_h\times d_t}$, and $\mathbf{w}_{\mathrm{pool}} \in \mathbb{R}^{d_h}$. Its raw query field is

$$a_i^{\mathrm{pool}}(F,x) = \mathbf{w}_{\mathrm{pool}}^{\top}\,\mathrm{GELU}(W_f\mathbf{f}_i + W_x\mathbf{x}).$$

Token–point head. The token–point head uses normalized non-padding tokens $T(x) \in \mathbb{R}^{n_t\times d_t}$ from the final text-transformer state:

$$Q = FW_q^{\top}, \qquad K = T(x)W_k^{\top}, \qquad V = T(x)W_v^{\top},$$

where $W_q \in \mathbb{R}^{d_h \times d}$ and $W_k, W_v \in \mathbb{R}^{d_h \times d_t}$. Thus $Q, U \in \mathbb{R}^{N \times d_h}$ and $K, V \in \mathbb{R}^{n_t \times d_h}$. The attended representation is

$$U = Q + \mathrm{Dropout}(\mathrm{MHA}(\mathrm{LN}(Q), \mathrm{LN}(K), V)) .$$

Here MHA is the standard eight-head scaled dot-product attention block, including its learned attention and output projections, and LN denotes layer normalization. The scalar query field is

$$a_i^{\mathrm{token}}(F, x) = \mathbf{w}_{\mathrm{token}}^{\top} \mathrm{FFN}(\mathrm{LN}([\mathbf{f}_i \| \mathbf{u}_i])) ,$$

where $\mathbf{u}_i$ is row i of U and $\mathbf{w}_{\mathrm{token}} \in \mathbb{R}^{d_h}$. The two-layer GELU FFN maps $\mathbb{R}^{d+d_h} \to \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}$. All dropout modules in this head use rate 0.1. Padding tokens are masked, and no token pooling follows the attention block.

Task loss. For a valid query with logit ℓ and target y, every controlled variant minimizes

$$\mathcal{L}_{\mathrm{task}} = \mathcal{L}_{\mathrm{BCE}}(\ell, y) + \lambda_{\mathrm{Dice}} \mathcal{L}_{\mathrm{Dice}}(\ell, y),$$

where

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2 \sum_i p_i y_i + \epsilon_D}{\sum_i p_i + \sum_i y_i + \epsilon_D}, \qquad p_i = \mathrm{sigmoid}(\ell_i).$$

We use $\lambda_{\mathrm{Dice}} = 1$ and $\epsilon_D = 10^{-6}$ for every variant. Absent prompts are excluded from this task loss.

B Experimental protocol and placement controls

B.1 Data isolation and statistical reporting

The official 3DCoMPaT200 training partition is divided by a fixed category-stratified manifest into $\mathcal{D}_{\mathrm{train}}$, $\mathcal{D}_{\mathrm{design}}$, and $\mathcal{D}_{\mathrm{confirm}}$. All trainable parameters are fit on $\mathcal{D}_{\mathrm{train}}$. The design split is used for checkpoint selection, the graph/filter configuration, and placement-control decisions. The counterfactual regressions are evaluated on $\mathcal{D}_{\mathrm{confirm}}$ after the model, prompt manifest, and analysis choices have been fixed. The confirmation split is excluded from parameter fitting and model selection. The segmentation metrics in the controlled result tables are evaluated on the named benchmark test sets, not on $\mathcal{D}_{\mathrm{confirm}}$.

Direct is the matched identity-placement control, not an external baseline. Direct and GSSR use the same pooled-prompt head and PTv3 backbone architectures, task loss, optimization schedule, and five paired training seed identities, but their parameters are optimized independently. For any metric whose seed-level outcomes are retained, the primary analysis unit is a training seed and paired differences are

$$d_s = M_s^{\mathrm{GSSR}} - M_s^{\mathrm{Direct}}.$$

Paired intervals, where shown, are two-sided 95% Student-t intervals over the seed-level differences. For the primary 3DCoMPaT-Coarse endpoint, Table 3 reports mean $\pm$ seed SD for Direct and GSSR across five paired training seeds, together with the paired mean difference and its 95% confidence interval. This interval is computed from the five seed-wise differences d_s using a Student-t distribution with four degrees of freedom. The remaining pooled-prompt benchmark mIoU and cross-dataset macro rows are reported as aggregate means without seed SDs or paired intervals. The same paired calculation remains available for the three token–point seeds. Object, query, and prompt-pair rows are not treated as independent training runs. With five seed pairs, a two-sided exact sign-flip test cannot attain $p < 0.05$; with three pairs, its minimum attainable two-sided value is 0.25. All retained intervals are therefore interpreted as descriptive effect estimates rather than confirmatory significance tests.

Valid-query metrics are first averaged within each object and then macro-averaged over objects. The primary endpoint is 3DCoMPaT-Coarse mIoU under reference orientation and object-aware prompts. The transfer macro is the unweighted mean of ShapeNet-Part and PartNet-E mIoU, and the four-dataset macro is the unweighted mean of the four reported dataset summaries. The separately retrained placement panel uses three paired seeds and is interpreted within that panel. Its 2.6-point Direct/GSSR mIoU contrast and the 3.37-point main-study contrast are reported separately under their respective protocols and are not pooled for inference.

B.2 Boundary and thin-part metrics

For a binary mask z, a union-symmetrized Euclidean k_b-NN metric graph defines the interior boundary

$$\mathcal{B}_{k_b}(z) = \{i : z_i = 1,\ \exists j \in \mathcal{N}_{k_b}(i) \text{ with } z_j = 0\}.$$

Strict BIoU_{16} is the set IoU between $\mathcal{B}_{16}(\mathbf{1}[p \geq 0.5])$ and $\mathcal{B}_{16}(y)$. Tolerant BF1 matches predicted and ground-truth boundary points within

$$\|\mathbf{p}_i - \mathbf{p}_j\|_2 \leq t\widetilde{d},$$

where $\widetilde{d}$ is the object's median 16-NN spacing; the primary setting uses $t = 1$. Specifically, precision is the fraction of predicted boundary points having at least one ground-truth boundary point within the tolerance, and recall reverses these roles. Let P and R denote this precision and recall. We report

$$\mathrm{BF1} = \frac{2PR}{P + R},$$

with BF1 set to zero when $P + R = 0$. This is a symmetric existence-based many-to-one tolerance rule rather than a one-to-one assignment. Exactly one empty boundary receives zero, whereas two empty boundaries are undefined. Absent queries are excluded from boundary metrics and evaluated separately by false-positive rate.

Boundary eligibility is fixed from ground truth: a valid query is eligible when its ground-truth boundary is nonempty. On this fixed set, an empty predicted boundary receives BIoU and BF1 equal to zero. Queries with an empty ground-truth boundary are excluded identically for every method. The retained aggregate record does not expose eligible, predicted-empty, and excluded counts by seed; this prevents a separate coverage audit but does not alter the fixed eligibility rule. The thin-part stratum is the top design-frozen quartile of $|\mathcal{B}_{16}(y)|/(|y| + 10^{-8})$.

B.3 Placement-control definitions

Write $b = b_\phi(F)$, $r = a_\theta(F, x) - a_\theta(F, x_\emptyset)$, and $G = I_N - H$. All low-, high-, and full-logit controls use the same $L = L_0$, η, and m as GSSR. The response

$$g(\lambda) = 1 - (1 - \eta\lambda)^m$$

is monotonically increasing on $\lambda \in [0, 2]$, and G removes the graph nullspace. For each connected component C, let

$$u_C = \frac{D_0^{1/2}\mathbf{1}_C}{\left\|D_0^{1/2}\mathbf{1}_C\right\|_2}, \qquad P_{\ker L} = \sum_C u_C u_C^\top.$$

Thus $P_{\ker L}$ is the orthogonal projector onto the graph nullspace, $r_0 = P_{\ker L} r$, and $r_\perp = (I_N - P_{\ker L})r$. The independently retrained controls use

$$\begin{aligned}
\ell_{\mathrm{Direct}} &= b + r, \\
\ell_{\mathrm{ResidualLow}} &= b + Hr, \\
\ell_{\mathrm{ResidualHigh}} &= b + P_{\ker L} r + G(I_N - P_{\ker L})r, \\
\ell_{\mathrm{FullLow}} &= H(b + r).
\end{aligned}$$

The Laplacian-regularized control keeps $\ell = b + r$ and optimizes

$$\mathcal{L}_{\mathrm{task}} + \lambda_L \frac{r_\perp^\top L r_\perp}{\|r_\perp\|_2^2 + 10^{-6}},$$

where $\lambda_L \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ is selected on the design split by object-macro mIoU, then BF1, then the smaller value. Each alternative is trained independently under the same three-seed control protocol.

The frozen intervention starts from each of the five Direct checkpoints and reuses its b and r without retraining. For an object-ID-seeded permutation Π, define $L^\pi = \Pi^\top L \Pi$ and $S^\pi = (I_N - \eta L^\pi)^m$. For

$$\begin{aligned}
O \in \{&I_N - P_{\ker L}, \\
&G(I_N - P_{\ker L}), \\
&(I_N - P_{\ker L}) S^\pi (I_N - P_{\ker L})\}.
\end{aligned}$$

Table 5: Additional held-out diagnostic outcomes. Values are descriptive means with 95% intervals under the aggregation in Appendix A.1.

Head	$\mathcal{E}_{L_0}(\Delta_{\mathrm{inv}})$	Absent FPR (%)
Pooled prompt (5 seeds)	$0.31\,[0.27, 0.35]$	$8.4\,[7.2, 9.6]$
Token–point (3 seeds)	$0.34\,[0.29, 0.39]$	$9.1\,[7.8, 10.5]$

Table 6: Three-seed token–point Direct/GSSR replication. Direct is the identity-placement control; C and F denote 3DCoMPaT-Coarse and 3DCoMPaT-Fine. Scores are object-macro percentages reported as mean ± seed SD. The macro difference includes its paired descriptive 95% Student-t interval over the three seed-level differences. BF1 is measured on 3DCoMPaT-Coarse at $t = 1$.

Method	C	F	ShapeNet-Part	PartNet-E	Macro	ΔMacro [95% CI]	BF1
Direct	47.2 ± 0.7	24.3 ± 0.7	35.9 ± 0.6	17.4 ± 0.6	31.2 ± 0.5	–	40.9 ± 0.9
GSSR	$\mathbf{48.2 \pm 0.6}$	$\mathbf{25.3 \pm 0.6}$	$\mathbf{36.7 \pm 0.5}$	$\mathbf{18.0 \pm 0.5}$	$\mathbf{32.1 \pm 0.4}$	$\mathbf{+0.9\,[0.2, 1.6]}$	$\mathbf{43.0 \pm 0.8}$

the norm-matching scale, computed separately for every seed–object–query tuple, is

$$s_O = \frac{\|Hr_\perp\|_2}{\|Or_\perp\|_2}.$$

The primary operator rows are, respectively,

$$\begin{aligned} \ell_{\mathrm{scalar}} &= b + r_0 + s_I r_\perp, \\ \ell_{\mathrm{high}} &= b + r_0 + s_G G r_\perp, \\ \ell_{\mathrm{shuffle}} &= b + r_0 + s_S (I_N - P_{\ker L}) S^\pi r_\perp. \end{aligned}$$

Full-logit diffusion uses $H(b + r)$. A seed–object–query tuple enters the reported common set only when every norm-matching denominator exceeds $10^{-6}\sqrt{N}$ and every scale is at most 10. The same intersection is used for all displayed operator rows. Because the retained aggregates do not include the excluded and eligible tuple counts, the frozen comparison is interpreted conditionally on this retained operator-valid common set.

C Additional controlled results

C.1 Additional diagnostic outcomes

Table 5 reports the mean normalized energy of the prompt-equivalent logit difference and the complementary absent-query false-positive rate. These outcomes accompany the regression coefficients in the main paper; absent queries are not used to define the paraphrase inconsistency target.

C.2 Token–point replication

Table 6 evaluates the same matched identity control/GSSR comparison with the token–point head. This replication uses three paired training seeds and is interpreted separately from the five-seed pooled-prompt comparison in the main paper.

GSSR has a $0.9\,[0.2, 1.6]$-point estimated mean increase in token–point macro mIoU and a 2.1-point higher observed mean BF1, providing a directionally consistent replication with a second fusion head.

C.3 Efficiency details

Table 7 reports the complete efficiency record under the timing protocol summarized in the main paper.

Table 7: Efficiency at $N = 5{,}000$. Timings use an RTX 4090 with AMP, 20 warm-up iterations, and 200 timed iterations; they include graph construction, CUDA synchronization, and the model forward pass, and exclude host-to-device transfer. Throughput is measured with batch size eight. GPU-hours are per 200-epoch training run.

Method	Trainable params	Batch-1	Infer. memory	Objects/s	GPU-hours
Direct	46.7M	69 ms	8.7 GB	103	75
GSSR	46.7M	73 ms	9.0 GB	98	75

Table 8: Complete external benchmark comparison in object-macro mIoU (%). Object-aware and Part-only prompts are {`part`} `of a` {`object`} and {`part`}, respectively. Published-baseline mIoU entries are taken from Table 1 of CoSMo3D under the same benchmark settings, whereas the GSSR entries report our five-seed means. Because run-level outcomes are unavailable for the published baselines, cross-method differences are descriptive rather than paired effect estimates.

Benchmark	Method	Reference orientation		Rotated	
		Object-aware	Part-only	Object-aware	Part-only
3DCoMPaT-Coarse	PointCLIP V2	14.09	14.16	13.17	13.39
	PartSLIP++	6.12	17.62	5.89	15.82
	Find3D	31.72	25.36	32.47	26.23
	CoSMo3D	47.51	54.52	47.74	54.55
	GSSR	**51.17**	**58.34**	**49.09**	**58.51**
3DCoMPaT-Fine	PointCLIP V2	7.25	7.18	7.43	7.20
	PartSLIP++	3.79	5.68	3.41	4.76
	Find3D	12.69	12.57	12.73	12.79
	CoSMo3D	24.48	31.29	24.46	30.97
	GSSR	**28.66**	**35.47**	**28.68**	**35.94**
ShapeNet-Part	OpenMask3D	8.94	10.37	6.75	14.56
	PointCLIP V2	16.91	20.22	16.88	18.19
	Find3D	28.39	24.09	29.64	23.71
	CoSMo3D	36.16	33.31	34.20	32.84
	GSSR	**38.71**	**37.70**	**36.87**	**35.82**
PartNet-E	OpenMask3D	12.54	11.24	11.93	11.67
	PointCLIP V2	11.28	9.70	10.32	10.22
	Find3D	16.86	16.38	17.62	17.16
	CoSMo3D	17.59	17.19	18.48	18.17
	GSSR	**19.52**	**19.38**	**19.87**	**20.39**

D EXTENDED EXTERNAL BENCHMARK COMPARISON

Tables 8 and 9 expand the published-method comparison in the main paper to both prompt forms and orientations. The published external baselines are PointCLIP V2, PartSLIP++, OpenMask3D, Find3D, and CoSMo3D. Published methods retain their own pipelines, so cross-method differences are descriptive. The GSSR rows are the same five-seed means used in the matched Direct/GSSR study; the paper does not maintain a second GSSR benchmark record.

Table 9: Object-macro BIoU_{16} (%) for the external benchmark comparison. Methods without retained point-level boundary results are omitted. Defined-case counts are unavailable; the values are therefore descriptive and are not formatted as a cross-method ranking.

Benchmark	Method	Reference orientation		Rotated	
		Object-aware	Part-only	Object-aware	Part-only
3DCoMPaT-Coarse	Find3D	2.86	1.18	2.81	1.97
	CoSMo3D	3.59	4.37	3.72	4.43
	GSSR	16.11	19.31	16.27	19.25
3DCoMPaT-Fine	Find3D	2.31	2.33	2.36	2.37
	CoSMo3D	3.37	5.19	3.56	5.45
	GSSR	10.81	14.67	11.78	15.67
ShapeNet-Part	Find3D	4.42	3.96	4.58	4.01
	CoSMo3D	4.85	5.11	4.68	5.04
	GSSR	16.30	15.75	16.24	14.71
PartNet-E	Find3D	0.18	0.16	0.19	0.15
	CoSMo3D	0.36	0.34	0.33	0.33
	GSSR	1.18	1.17	0.93	0.95

GSSR has the largest reported mIoU mean among the available method entries in all 16 settings. BIoU_{16} is reported only as descriptive context because retained defined-case counts do not support a complete cross-method coverage audit.