TRACING CHAINS OF EVIDENCE: CROSS-VIEW TEST-TIME ADAPTATION FOR POINT CLOUD PART SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Point cloud part segmentation increasingly lifts multi-view predictions from 2D foundation models into 3D, reducing the need for costly point-level annotations. To assess the robustness of this paradigm, we establish, to the best of our knowledge, the first distribution-shift benchmark for 2D-to-3D part segmentation, spanning corruption, category, and sim-to-real shifts. Our evaluation reveals substantial performance degradation and exposes an inherent evidence-coupling challenge: multi-view predictions linked through shared 3D geometry can be complementary, ambiguous, or conflicting, hindering their reliable integration. To address this challenge, we propose **T**est-time **R**eliable **A**daptation via **C**hain-of-**E**vidence (**TRACE**), a novel TTA framework that recasts adaptation as discovering and reinforcing structural consensus within entangled evidence. Specifically, TRACE constructs a weighted evidence graph that jointly models semantic compatibility and 3D geometric consistency, organizing compatible cross-view predictions into evidence chains while routing ambiguous evidence to an explicit uncertainty state. To avoid the exponential cost of enumerating these chains, we derive an exact composition rule through mathematical induction that aggregates their support in closed form with linear complexity. Together, these designs move TTA beyond confidence-driven refinement of isolated predictions toward uncertainty-aware reasoning over structured cross-view consensus, enabling robust integration of complementary evidence. Extensive experiments across diverse domain shifts and base models demonstrate that TRACE achieves state-of-the-art performance, with an average improvement of **+2.6%** mIoU. Code will be publicly available.

1 INTRODUCTION

Point cloud part segmentation aims to assign fine-grained semantic labels to individual points, enabling precise structural understanding for applications such as 3D content creation, shape analysis and synthesis, and robotic manipulation (Mo et al., 2019; Geng et al., 2023; Liu et al., 2024a; Chen et al., 2025). However, collecting dense point-level annotations for 3D data is expensive and difficult to scale, particularly under fine-grained or open-vocabulary part taxonomies (Li et al., 2018; Wang et al., 2018; Ma et al., 2025). Recent studies therefore explore leveraging powerful 2D foundation models: they render a point cloud from multiple viewpoints, obtain part predictions in the image space, and subsequently lift these predictions back to 3D space (Zhou et al., 2023; Kim & Sung, 2024; Deng et al., 2026). This 2D-to-3D lifting paradigm converts the key challenge from learning 3D semantics from scratch to consolidating multi-view 2D predictions in 3D space.

Despite its progress in low-data regimes, the lifting pipeline introduces a reliability bottleneck: the 3D outputs inevitably inherit errors from 2D prediction, which can be sensitive to rendering artifacts, geometric perturbations, and unseen object semantics (Genova et al., 2021; Ren et al., 2022). To systematically study this challenge, we establish, to the best of our knowledge, the first distribution-shift benchmark for 2D-to-3D part segmentation. It covers three practical scenarios: (1) corruption shifts, arising from degraded geometry; (2) category shifts, involving novel object categories; and (3) sim-real shifts, which evaluate transfer from synthetic CAD models to real-world scans. As shown in Fig. 1(a), extensive experiments show substantial performance degradation across these shifts, revealing that direct lifting alone is insufficient for robust 3D part segmentation.

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

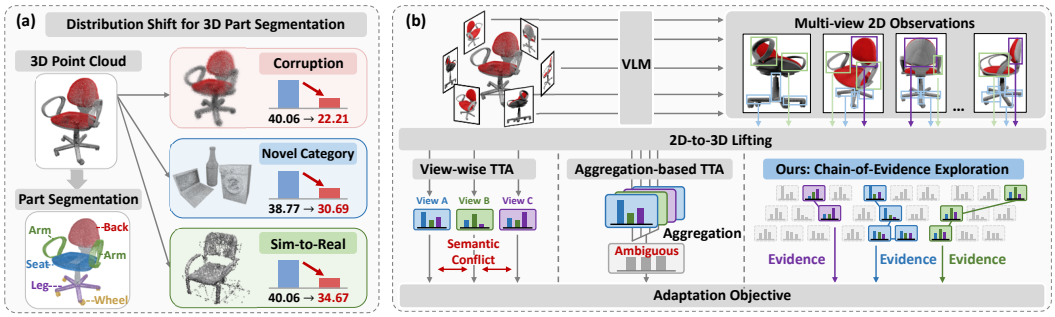

Figure 1: (a) Impact of distribution shifts on performance. (b) Different adaptation formulation in the lifting pipeline: view-wise TTA adapts each view independently; aggregation-based TTA adapts using fused prediction; and TRACE adapts by tracing structured evidence consensus across views.

To mitigate such distribution shifts with minimal overhead, test-time adaptation (TTA) (Wang et al., 2021; Liang et al., 2025) has emerged as a critical paradigm, enabling pretrained models to adjust to target distributions through online updates on unlabeled samples. Most TTA methods construct self-supervision from the model's own predictions, regarding high-confidence predictions as reliable evidence for online adaptation (Niu et al., 2022; 2023). Typical strategies include entropy minimization that sharpens prediction distributions (Zhang et al., 2022), pseudo-labeling that reinforces selected classes (Chen et al., 2022), and uncertainty-aware adaptation that filters unreliable samples (Hu et al., 2025a). From this perspective, TTA can be viewed as increasing the model's confidence on target evidence while avoiding excessive reinforcement of noisy predictions.

While TTA has achieved considerable success in 2D tasks, directly applying it to the lifting pipeline is challenging due to an inherent **evidence-coupling dilemma**. In multi-view 3D perception, 3D points are observed from multiple viewpoints whose predictions are linked by the geometric structure. These observations constitute correlated evidence that can be complementary, ambiguous, or even contradictory, rather than independent adaptation signals. This coupling exposes a fundamental limitation: view-wise TTA ignores cross-view semantic conflict and may drive the model toward incompatible part predictions, whereas aggregation-based TTA can mix reliable evidence with ambiguous cues and reinforce unreliable supervision. Both strategies fail to capture the underlying evidence structure, hindering the reliable integration of complementary multi-view evidence.

To address these challenges, we propose a novel TTA framework "**T**est-time **R**eliable **A**daptation via **C**hain-of-**E**vidence (**TRACE**)", which recasts online adaptation as discovering and reinforcing structural consensus within entangled evidence, as shown in Fig. 1(b). TRACE progressively links compatible predictions across views into chains, tracing how fragmented evidence collectively supports a coherent hypothesis. Beyond evidence accumulation, these chains captures both semantic agreement and geometric consistency, providing a unified criterion for assessing reliability. By consolidating category-aware weights along discovered chains, it identifies trustworthy hypotheses and adaptively reinforces their supporting evidence, enabling coordinated enhancement across views.

Specifically, TRACE represents multi-view predictions as a weighted evidence graph, where vertices denote semantic states, edges encode semantic compatibility, and vertex weights measure evidence strength. To ground compatibility in 3D space, TRACE calibrates the graph using geometric consensus derived from a 3D affinity matrix, transferring geometrically unsupported evidence to an explicit uncertainty state. Building on the calibrated graph, each complete evidence chain emerges as a maximum clique, whose accumulated weight quantifies the joint support for a category hypothesis. However, computing this support by enumerating all candidate cliques incurs a combinatorial cost that grows exponentially with the graph size. To bypass this lengthy process, we leverage mathematical induction to derive a composition rule that all clique weights can be exactly marginalized in closed form. The resulting support parameterizes an uncertainty-aware Dirichlet distribution that gates online updates, reinforcing reliable evidence while suppressing uncertain supervision. Together, these designs shift the basis of adaptation from isolated prediction confidence to structured evidence consensus, providing a unified mechanism for integrating complementary evidence.

We evaluate the effectiveness and generalizability of our method across diverse shift types and base models, achieving state-of-the-art results with an average improvement of **+2.6%** mIoU. Furthermore, we showcase its practicality and reliability in a phone-scanning device setup, where it produces clear refinements in segmentation quality over existing methods.

Contributions. 1) To the best of our knowledge, we establish the first benchmark to evaluate the OOD robustness of 2D-to-3D lifting paradigms, providing a comprehensive testbed under diverse and realistic deployment conditions. 2) We identify the *evidence-coupling dilemma* in multi-view TTA and propose TRACE, which reformulates online adaptation as tracing and reinforcing compatible chains within an evidence graph. 3) We derive an exact composition rule that marginalizes support over all evidence chains without explicit enumeration, reducing the computational complexity from exponential to linear. 4) Extensive experiments demonstrate the superiority of TRACE, validating the effectiveness of structured evidence consensus for multi-view adaptation.

2 Related Work

2.1 Point Cloud Part Segmentation

The early development of point cloud part segmentation was largely driven by supervised learning under closed label settings from datasets such as ShapeNet-Part (Yi et al., 2016) and PartNet (Mo et al., 2019). Representative methods employ geometric encoders, including PointNet/PointNet++(Qi et al., 2017a;b) and the Point Transformer family (Zhao et al., 2021; Wu et al., 2022), to capture global shape representations with local structural context. However, dense point-level annotation is costly, limiting scalability to broader object categories and part vocabularies.

Recent advances in 2D foundation models offer a promising means of transferring rich semantic priors to 3D, reducing reliance on native 3D annotations (Zhou et al., 2023; Deng et al., 2026). This 2D-to-3D pipeline renders a 3D shape from multiple viewpoints, predict semantics in image space, and lifts them to 3D through backprojection and cross-view fusion. For example, PartSLIP (Liu et al., 2023) applies GLIP (Li et al., 2022) to rendered views and consolidates part detections through multi-view voting, whereas PartSTAD (Kim & Sung, 2024) refines detected regions with SAM (Kirillov et al., 2023) and learns a weighting network for task-aware fusion. These works focus on improving evidence generation and fusion under standard settings, leaving robustness to real-world distribution shifts largely unexplored. Our work fills this gap by establishing a distribution-shift benchmark and improving lifting robustness through TTA.

2.2 Test-Time Adaptation

Test-time adaptation (TTA) adapts a pretrained model to unlabeled target data during inference without additional annotation (Lim et al., 2023; Yuan et al., 2023). Depending on whether the source training process is modified, it can be broadly categorized into test-time training (Hakim et al., 2023; Liu et al., 2024b; Durasov et al., 2025) and fully test-time adaptation (Boudiaf et al., 2022; Tsai et al., 2024; Zhou et al., 2025). We focus on the latter for its applicability to off-the-shelf models. Fully TTA typically extract reliable evidence from target predictions and reinforce it while suppressing noisy updates (Lee et al., 2024; Sun et al., 2024). For example, SAR (Niu et al., 2023) treats low-entropy predictions as reliable evidence and reinforces them through sharpness-aware minimization. ReCAP (Hu et al., 2025a) extends reliability estimation to local regions in the embedding space.

While effective across diverse vision tasks (Shin et al., 2022; Lin et al., 2024; Hu et al., 2026), existing methods assess evidence independently and overlook cross-view dependencies in multi-view lifting. Their direct application yields two strategies. *View-wise adaptation* adapts each view independently, ignoring 3D correspondences and reinforcing inconsistent part prediction. *Aggregation-based adaptation* adapts on fused predictions, obscuring evidence provenance and reinforcing ambiguous agreement. In this work, TRACE organizes cross-view predictions into a weighted evidence graph and adapts only from semantically compatible and geometrically supported consensus.

3 Preliminary

Task Definition. We consider point cloud part segmentation in a streaming setting. Given a segmentation network f_θ pre-trained on a source dataset, the model sequentially processes a stream of unlabeled target point clouds $\mathcal{X} = \{X_1, \ldots, X_N\}$. Each point cloud $X = \{\mathbf{p}_n\}_{n=1}^{\mathcal{N}} \in \mathbb{R}^{\mathcal{N}\times D}$ contains $\mathcal{N}$ points, where $\mathbf{p}_n$ denotes the D-dimensional representation of the n-th point. For an incoming point cloud, the model produces a semantic part label for each point:

$$\mathbf{z}_i = f_\theta(X_i), \quad \hat{y}_{i,n} = \arg\max_{c\in\{1,\ldots,C\}} \mathbf{z}_{i,n}^c, \tag{1}$$

where C denotes the number of part categories specified by the user-provided text prompts, $\mathbf{z}_{i,n} \in \mathbb{R}^C$ denotes the class-wise probability vector, and $\hat{y}_{i,n}$ is its predicted part label.

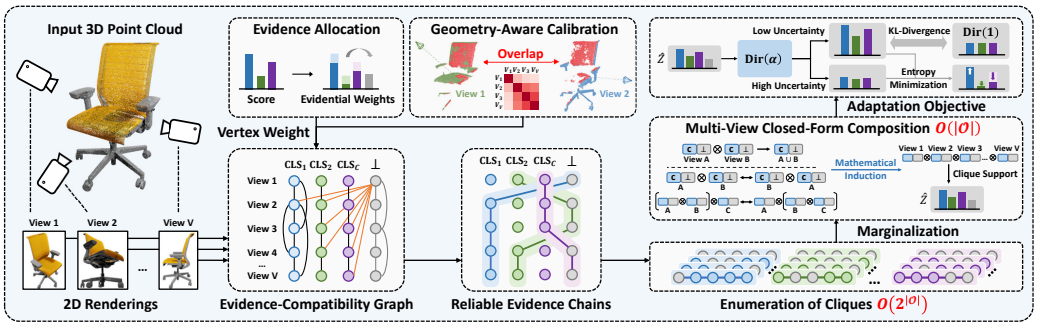

Figure 2: Overview of our TRACE framework. In Sec. 4.1, TRACE organizes multi-view 2D evidence into a compatibility graph, where reliable chains correspond to maximum cliques. To avoid exponential clique enumeration, in Sec. 4.2, TRACE derives a binary composition rule and its closed-form extension, driving an uncertainty-gated Dirichlet objective for reliable adaptation.

2D-to-3D Lifting Framework. Recent works formulate this task as a multi-view perception problem via the 2D-to-3D lifting framework (Liu et al., 2023; Kim & Sung, 2024). It first renders V images $\{I^v\}_{v=1}^V$ from predefined viewpoints. For each view, a 2D foundation model f takes the rendered image I^v and part-category prompts T, producing a set of region-level evidence:

$$f(I^v, T) = \{e_i^v\}_{i=1}^{M_v}, \quad e_i^v = (r_i^v, \mathbf{s}_i^v), \tag{2}$$

where M_v denotes the number of detected regions. Each evidence term e_i^v contains a localized region r_i^v (box or mask) with class-wise scores $\mathbf{s}_i^v \in \mathbb{R}^C$. The 2D evidence is then lifted to 3D using the known camera projections. For each 3D point, we gather all evidence whose localized regions cover the point projected location, and compute the lifted prediction as a weighted sum:

$$\mathcal{O}_n = \{v \mid \mathbf{p}_n \text{ is visible in } r^v\}, \quad \mathbf{z}_n = \sum_{v \in \mathcal{O}_n} w_n^v \, \mathrm{softmax}(\mathbf{s}^v), \tag{3}$$

where $\mathcal{O}_n$ denotes visible views and w is determined by a learned module or non-parametric voting.

4 Methodology

In this paper, we propose TRACE (**T**est-time **R**eliable **A**daptation via **C**hain-of-**E**vidence), a novel TTA framework for 2D-to-3D part segmentation. TRACE incorporates two core designs: chain-of-evidence construction with semantic-geometric compatibility graph (Sec. 4.1), and closed-form evidence reinforcement via composition rule (Sec. 4.2). An overview is illustrated in Fig. 2.

4.1 Turning Multi-View Evidence into Reliable Chains

Existing TTA methods typically derive adaptation signals from individual predictions (Wang et al., 2021; Niu et al., 2022), whereas evidence for a 3D point is jointly distributed across geometrically coupled views. We therefore organize multi-view predictions into *chains of evidence* that are semantically compatible and geometrically grounded across visible views. To this end, we build a semantic-geometric compatibility graph over view-state evidence, where each chain corresponds to a global explanation of how multi-view evidence jointly supports a coherent 3D prediction.

Semantic-Compatibility Graph. Given a point $\mathbf{p}$ observed in views $\mathcal{O}$, we organize its multi-view predictions as a weighted graph $\mathcal{G}=(\mathcal{V}, \mathcal{E} \mid \mathcal{W})$, where $\mathcal{V}$, $\mathcal{E}$, and $\mathcal{W}$ denote the vertices, edges, and vertex weights. For each view, we augment the C semantic states with an uncertainty state, allowing ambiguous observations to remain in the graph without being forced into an unreliable label:

$$\mathcal{V} = \{(v, s) \mid v \in \mathcal{O},\ s \in \mathcal{S}\}, \quad \mathcal{S} = \{1, \ldots, C, \perp\}, \tag{4}$$

where $\perp$ denotes uncertainty. We assign each vertex a normalized weight by competing class-specific score with a fixed uncertainty reserve. Given the class-wise logits $\mathbf{z}^v$, we convert them into non-negative evidence $\mathbf{e}^v = \mathrm{softplus}(\mathbf{z}^v)$ and normalize it with the uncertainty budget:

$$w(v, s) = \frac{e^{v,s}}{\sum_{i=1}^{C}(e^{v,i} + m)}, \ s \in \{1, \ldots, C\}; \quad w(v, \perp) = \frac{Cm}{\sum_{i=1}^{C}(e^{v,i} + m)}. \tag{5}$$

where $m > 0$ controls the uncertainty budget. This normalization assigns mass according to semantic evidence while favoring uncertainty when the evidence is weak.

With vertex weights specifying evidence strength, edges encode semantic compatibility across views. For two vertices (v, s) and (v', s') with $v \neq v'$, we add an undirected edge if their states satisfy one of three rules: 1) both states indicate the same class, 2) both states are uncertain, or 3) one state is semantic and the other is uncertain. These rules connect mutually compatible evidence while excluding cross-view conflicts, forming the graph basis for chain-of-evidence construction.

Geometry-Aware Evidence Calibration. The above graph structure depicts the state coexistence at the semantic level, but semantic agreement alone does not ensure geometric reliability. A compatible prediction may still be hallucinated if the projected evidence comes from different 3D regions due to projection ambiguity or occlusion. We therefore assess geometric support by comparing the visible 3D point sets $\mathcal{Q}^v$ covered by the projected 2D regions of different views. For each pair, we define the element of the geometric affinity matrix G as the intersection-over-union of their support sets:

$$G_{v,v'} = \frac{|\mathcal{Q}^v \cap \mathcal{Q}^{v'}|}{|\mathcal{Q}^v \cup \mathcal{Q}^{v'}|}, \quad v, v' \in \mathcal{O}. \tag{6}$$

A large $G_{v,v'}$ indicates that two views share evidence from same local 3D support, whereas a small value suggests mismatch or ambiguous projection. Rather than estimating view reliability independently, we derive it from the global geometric consensus. Specifically, we compute the reliability vector $\boldsymbol{\rho} = [\rho^v]$ as the dominant eigenvector and normalize it by the maximum entry:

$$G\boldsymbol{\rho} = \lambda_{\max}\boldsymbol{\rho}, \quad \rho^v \leftarrow \frac{\rho^v}{\max_{v' \in \mathcal{O}} \rho^{v'}}. \tag{7}$$

This vector assigns a high score to a view only when it overlaps with other reliable views, while downweighting views whose support is isolated. We then use it to calibrate the graph weights in a mass-preserving manner: semantic evidence is scaled by its geometric reliability, and the suppressed mass is transferred to the uncertainty state:

$$\bar{w}(v, s) = \rho^v w(v, s),\ s \in 1, \ldots, C; \quad \bar{w}(v, \perp) = w(v, \perp) + (1 - \rho^v) \sum_{c=1}^{C} w(v, c). \tag{8}$$

This calibration converts geometric consensus into an evidence transfer mechanism, grounding the evidence chain by dynamically suppressing hallucinated semantic confidence.

Maximum-Clique Evidence Consolidation. Since states from the same view are disconnected, a clique selects at most one state per view; since cross-view edges encode semantic compatibility, all selected states are mutually consistent. A complete evidence chain therefore corresponds to a maximum clique of size $|\mathcal{O}|$ Each maximum clique $\mathcal{K}$ induces a unique state $\ell(\mathcal{K}) \in \mathcal{S}$: if a clique contains explicit class states, semantic compatibility ensures that they all indicate the same class; otherwise, the all-uncertainty clique is assigned to $\perp$. We group maximum cliques by their induced states and accumulate their calibrated weight products to obtain hypothesis-level support:

$$Z_s = \sum_{\mathcal{K} \in \mathrm{MaxCli}(\mathcal{G})} \mathbb{I}\left[\ell(\mathcal{K}) = s\right] \prod_{(v,s') \in \mathcal{K}} \bar{w}(v, s'), \quad \hat{Z}_s = \frac{Z_s}{\sum_{s' \in \mathcal{S}} Z_{s'}}, \tag{9}$$

where $\mathrm{MaxCli}(\cdot)$ denotes the set of maximum cliques, and $Z_s / \hat{Z}_s$ are the unnormalized/normalized supports, respectively. In this way, evidence consolidation is performed over complete and globally compatible chains, rather than isolated view confidences or prematurely fused scores.

4.2 Chain-of-Evidence Reinforcement

While the evidence graph establishes a principled correspondence between reliable chains and maximum cliques, this expressive formulation introduces a computational challenge: directly computing all chain hypotheses requires summing over exponentially many cliques (Moon & Moser, 1965), i.e., $O(2^{|\mathcal{O}|})$, which is impractical for TTA. Leveraging the mathematical induction, we prove that the proposed clique consolidation admits an exact closed-form reduction, allowing all maximum-clique hypotheses associated with each state to be marginalized without explicit enumeration.

Closed-Form Clique Composition. Given the evidence graph $\mathcal{G}$, explicit clique enumeration is not necessary because the proposed compatibility structure admits a view-wise compositional form. We first expose this structure in the binary case, where the supports of two disjoint view subsets can be merged. This elementary rule serves as the induction base for the general multi-view composition.

Proposition 1 (**Binary Composition Rule**). *Let $\mathcal{A}$ and $\mathcal{B}$ be two disjoint subsets of valid views, and let $\mathbf{Z}^{\mathcal{A}}$ and $\mathbf{Z}^{\mathcal{B}}$ denote the hypothesis-level supports obtained from their induced subgraphs. The support over the union $\mathcal{A} \cup \mathcal{B}$ satisfies the following binary composition:*

$$Z_{\perp}^{\mathcal{A}\cup\mathcal{B}} = Z_{\perp}^{\mathcal{A}} Z_{\perp}^{\mathcal{B}},\ Z_{c}^{\mathcal{A}\cup\mathcal{B}} = Z_{c}^{\mathcal{A}} Z_{\perp}^{\mathcal{B}} + Z_{c}^{\mathcal{B}} Z_{\perp}^{\mathcal{A}} + Z_{c}^{\mathcal{A}} Z_{c}^{\mathcal{B}}. \tag{10}$$

Moreover, this rule is commutative and associative, i.e., $\mathbf{Z}^{\mathcal{A}\cup\mathcal{B}} = \mathbf{Z}^{\mathcal{B}\cup\mathcal{A}}$ and $\mathbf{Z}^{(\mathcal{A}\cup\mathcal{B})\cup\mathcal{C}} = \mathbf{Z}^{\mathcal{A}\cup(\mathcal{B}\cup\mathcal{C})}$.

Building on this atomic composition rule, we extend the consolidation to an arbitrary number of valid views. The resulting closed-form solution marginalizes the sum over all clique associated with each state, bypassing the lengthy enumeration process.

Proposition 2 (**Multi-View Closed-Form Composition**). *Let $\mathcal{G}$ be the evidence graph for point $\mathbf{p}$ with valid views $\mathcal{O}$, and let $\bar{w}_v(s) = \bar{w}(v, s)$ denote the calibrated weight of state s in view v. For each state $s \in \mathcal{S}$, the maximum-clique support admits the following equivalent closed form:*

$$\hat{Z}_s = \begin{cases} \dfrac{\prod_v \bar{w}_v(\perp)}{\sum_{c=1}^{C} \prod_v (\bar{w}_v(c) + \bar{w}_v(\perp)) - (C-1) \prod_v \bar{w}_v(\perp)}, s = \perp, \\ \dfrac{\prod_v (\bar{w}_v(s) + \bar{w}_v(\perp)) - \prod_v \bar{w}_v(\perp)}{\sum_{c=1}^{C} \prod_v (\bar{w}_v(c) + \bar{w}_v(\perp)) - (C-1) \prod_v \bar{w}_v(\perp)}, s \neq \perp, \end{cases} \tag{11}$$

where the denominator is the normalization constant over $\mathcal{S}$. The first branch represents the all-uncertainty chain. The second branch expands all compatible chains that contain class-s evidence

Notably, Proposition 2 converts clique consolidation into exact composition rule, preserving the full Chain-of-Evidence support without explicit clique enumeration. This reduces computation from $O(2^{|\mathcal{O}|})$ to $O(|\mathcal{O}|)$ per state, making structured evidence reinforcement practical for TTA.
The proof follows by induction over the number of views and is provided in Appendix A.

Dirichlet-based Adaptation Objective. The normalized chain support $\hat{Z}_s$ in Eq. 11 jointly characterizes the semantic preference and uncertainty of the consolidated multi-view evidence. To obtain an uncertainty-aware distribution over semantic predictions, we parameterize a Dirichlet distribution using the semantic supports $\{\hat{Z}_c\}_{c=1}^{C}$ and the uncertainty support $\hat{Z}_{\perp}$:

$$\boldsymbol{p} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\ \alpha_c = \frac{\hat{Z}_c + \hat{Z}_{\perp}}{\hat{Z}_{\perp}},\ c \in \{1, \ldots, C\}. \tag{12}$$

A small $\hat{Z}_{\perp}$ yields large concentration parameters, indicating that the semantic prediction is supported by reliable evidence chains. When $\hat{Z}_{\perp}$ is large, $\boldsymbol{\alpha}$ stays close to the no-evidence prior $\mathbf{1}$, preventing uncertain chains from producing over-confident supervision. This conversion turns discrete chain support into a differentiable prediction for online adaptation.

We then optimize the model with an uncertainty-gated adaptation objective. Let $c^{\star} = \arg\max_c \hat{Z}_c$ denote the semantic class with the strongest consolidated chain support. We reinforce its expected probability while encouraging the Dirichlet distribution to deviate from the no-evidence prior only when the consolidated uncertainty is sufficiently low:

$$\mathcal{L}_{\mathrm{TTA}} = -I[\hat{Z}_{\perp} \leq \tau] \left(\log \mathbb{E}[p_{c^{\star}}] + \lambda \cdot \mathbb{D}_{KL} \left(\mathrm{Dir}(\boldsymbol{\alpha}) \| \mathrm{Dir}(\mathbf{1}) \right) \right), \tag{13}$$

where $\mathbb{E}[\cdot]$ denotes expectation under $\mathrm{Dir}(\boldsymbol{\alpha})$, D_{KL} denotes the KL-divergence, τ is the uncertainty threshold, and λ controls the strength of evidence reinforcement. The compatibility graph removes semantically conflicting chains, while the uncertainty gate disables adaptation when the remaining evidence is dominated by the uncertainty state, as indicated by a large $\widehat{Z}_{\perp}$. Together, they ensure that prompt updates are driven only by reliable cross-view evidence.

4.3 Overall Procedure of TRACE

For each 3D point and visible view, TRACE selects the smallest valid region covering its projection, using its prediction as the view-specific semantic evidence $\mathbf{z}_n^v$ and its lifted point set as the geometric support $\mathcal{Q}_n^v$. TRACE then constructs a semantic-geometric compatibility graph to organize cross-view evidence via Eq. 5 and calibrate geometrically unreliable semantics via Eq. 8. The calibrated evidence is consolidated into state-wise supports through the closed-form chain composition in Eq. 11, which are further parameterized as an uncertainty-aware Dirichlet distribution. Following the efficient prompt tuning for VLM (Zhou et al., 2022a;b; Yao et al., 2023), TRACE optimizes the uncertainty-gated adaptation objective in Eq. 13 by updating only the learnable text prompts while keeping the perception backbone frozen.

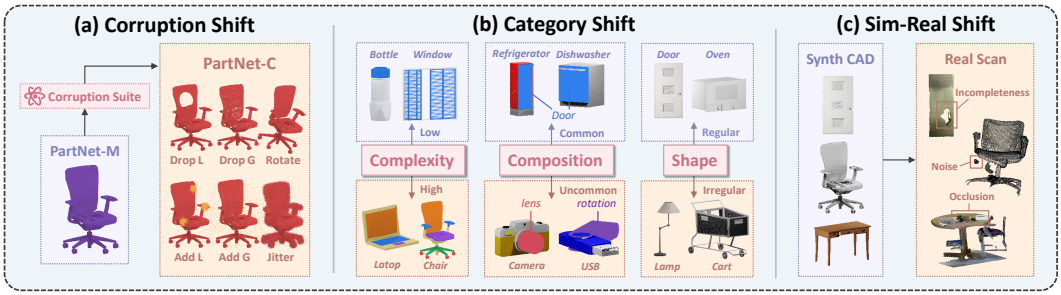

Figure 3: An illustration of three scenarios in our benchmark. (a) Diverse geometric degradations. (b) Variations in structural complexity, part composition, and shape regularity. (c) Transfers from synthetic CAD models to real scans with incompleteness, noise, and occlusion.

Table 1: Corruption-shift results on PartNet-C. We report mIoU (%) averaged over three independent runs. Within each block, the best and second-best are highlighted in **bold** and underline.

Model+Method	Severity Level 3							Severity Level 5						
	Drop-L	Drop-G	Add-L	Add-G	Rotate	Jitter	Avg.	Drop-L	Drop-G	Add-L	Add-G	Rotate	Jitter	Avg.
PartSLIP (Source)	32.03	31.27	24.49	24.68	30.63	10.98	25.68	25.12	24.76	19.27	19.66	24.30	8.72	20.31
• TENT (ICLR'21)	13.20	14.37	23.96	18.39	26.50	7.21	17.27	8.78	18.20	14.11	18.21	16.01	0.82	12.69
• SAR (ICLR'23)	15.98	18.98	23.59	20.72	21.69	6.45	17.90	13.91	16.52	16.58	14.69	19.35	5.83	14.48
• DUO (ICCV'25)	26.83	26.36	24.75	19.69	19.41	8.72	20.96	19.98	23.15	17.79	20.05	24.26	7.59	18.80
• ReCAP (ICML'25)	32.02	31.55	24.58	25.07	30.97	11.18	25.90	25.22	25.05	20.12	19.78	25.50	8.25	20.65
• ZeroSiam (ICLR'26)	32.86	32.38	25.22	25.73	29.79	11.45	26.23	26.10	24.40	20.21	18.98	25.14	8.68	20.59
• TRACE (Ours)	**35.08**	**34.86**	**28.92**	**29.73**	**33.44**	**12.32**	**29.06**	**28.20**	**27.68**	**21.23**	**22.26**	**26.58**	**9.60**	**22.59**
PartSTAD (Source)	36.48	36.26	34.49	32.97	32.45	13.02	30.95	28.33	27.96	20.32	20.28	26.79	9.59	22.21
• TENT (ICLR'21)	22.02	27.10	30.85	22.56	27.38	10.31	23.37	18.82	10.50	22.38	19.55	23.55	5.15	16.66
• SAR (ICLR'23)	23.46	29.85	30.12	22.69	28.14	12.59	24.48	20.15	14.78	17.34	17.92	23.48	7.60	16.88
• DUO (ICCV'25)	34.92	35.89	26.77	27.41	29.63	13.96	28.09	23.35	25.50	17.00	14.59	25.36	7.92	18.95
• ReCAP (ICML'25)	37.16	38.72	36.90	34.66	34.44	14.38	32.71	30.17	28.03	20.72	19.63	27.50	8.37	22.40
• ZeroSiam (ICLR'26)	37.07	38.57	37.36	35.01	34.19	14.92	32.85	28.80	27.96	20.08	21.25	28.88	8.08	22.51
• TRACE (Ours)	**40.95**	**40.78**	**39.68**	**37.37**	**37.98**	**16.86**	**35.60**	**33.08**	**30.74**	**21.38**	**24.79**	**31.40**	**10.03**	**25.24**

5 Experiments

5.1 Experimental Setup

We evaluate TRACE on our benchmark built on PartNet-Mobility (Xiang et al., 2020) and ScanObjectNN (Uy et al., 2019). For fair comparison, all methods share the same source model and 2D-to-3D lifting pipeline, following the evaluation protocols of prior work.

Benchmark Design. To systematically evaluate robustness, we construct, to the best of our knowledge, the first distribution-shift benchmark for 2D-to-3D part segmentation. As shown in Fig. 3, it covers three complementary shifts: geometric corruption, category variation and simulation-to-real transfer, respectively testing changes in point-cloud quality, semantic structure and acquisition conditions. Together, these settings provide a comprehensive testbed for evaluating TTA under diverse deployment scenarios. Detailed settings are provided in Appendix B.

Compared Methods. Based on two representative base models PartSLIP (Liu et al., 2023) and PartSTAD (Kim & Sung, 2024), we compare TRACE with several SOTA methods: TENT (Wang et al., 2021) minimizes prediction entropy on target data; SAR (Niu et al., 2023) combines reliable-sample selection with sharpness-aware entropy minimization; ReCAP (Hu et al., 2025a) performs region-level confidence optimization in feature space; DUO (Hu et al., 2025b) jointly models semantic and geometric uncertainty for 3D adaptation; and ZeroSiam (Chen et al., 2026) employs asymmetric Siamese learning with frozen and learnable branches.

Implementation Details. Following prior work (Liu et al., 2023; Kim & Sung, 2024), we render each object from 10 fixed viewpoints and equip the VLM with learnable soft prompts. During TTA, only the prompts are updated, while the pretrained VLM and 2D-to-3D lifting pipeline remain frozen. We optimize the prompts with AdamW using a learning rate of 0.07 and weight decay of 0.25. For the compared methods, we evaluate both *view-wise* and *aggregation-based* strategies. Since the latter generally performs better, we use it as the default setting in the main paper and report the complete results of both modes in Appendix C.1. For TRACE, we set the uncertainty budget to $m = 0.9$, the gate threshold to $\tau = 0.5$, and the weighting coefficient to $\lambda = 1.0$.

Table 2: Category-shift results on six adaptation tasks from PartNet-Mobility. We report mIoU averaged over three independent runs. L/H, C/U, and R/I denote low/high complexity, common/ uncommon composition, and regular/irregular shape. Time reports per-instance adaptation time.

Model+Method	Complexity			Composition			Shape			Avg.	Time(s)
	L→H	H→L	Avg.	C→U	U→C	Avg.	R→I	I→R	Avg.		
PartSLIP (Source)	25.91	20.34	23.13	29.04	15.91	22.48	30.74	16.42	23.58	23.06	6.56
• TENT (ICLR'21)	20.02	20.19	20.10	17.92	6.31	12.12	20.32	9.42	14.86	15.69	32.48
• SAR (ICLR'23)	26.39	20.15	23.27	25.32	4.22	14.76	22.75	8.92	15.84	17.96	29.91
• DUO (ICCV'25)	20.65	20.15	20.40	20.65	9.97	15.31	32.65	15.08	23.87	19.86	35.69
• ReCAP (ICML'25)	27.69	20.50	24.10	28.80	15.39	22.10	30.12	16.48	23.30	23.17	28.19
• ZeroSiam (ICLR'26)	27.29	19.89	23.59	29.74	16.48	23.11	24.01	17.22	20.62	22.44	27.41
• TRACE (Ours)	**30.28**	**21.54**	**25.91**	**30.12**	**17.98**	**24.05**	**37.49**	**17.56**	**27.53**	**25.83**	28.09
PartSTAD (Source)	31.71	26.59	29.15	31.92	29.19	30.56	37.95	26.80	32.38	30.69	6.98
• TENT (ICLR'21)	25.14	22.30	23.72	24.46	24.30	24.38	20.39	25.77	23.08	23.73	34.27
• SAR (ICLR'23)	18.14	24.51	21.33	32.61	27.86	30.24	22.76	19.16	20.96	24.18	33.25
• DUO (ICCV'25)	25.04	21.64	23.34	27.55	26.85	27.20	34.53	26.25	30.39	26.97	37.92
• ReCAP (ICML'25)	33.68	26.05	29.87	32.90	29.22	31.06	37.09	27.05	32.07	31.00	31.75
• ZeroSiam (ICLR'26)	35.50	26.45	30.98	29.42	31.35	30.39	38.00	27.91	32.96	31.44	29.81
• TRACE (Ours)	**39.80**	**29.03**	**34.42**	**37.55**	**32.70**	**35.13**	**41.39**	**29.34**	**35.37**	**34.97**	31.45

Table 3: Sim-real shift results on the ScanObjectNN benchmark. We report mIoU (%) averaged over three independent runs. Each column denotes an adaptation task for one target category.

Model+Method	Bed	Box	Cabinet	Chair	Desk	Display	Door	Shelf	Sink	Sofa	Table	Toilet	Avg.
PartSLIP (Source)	28.92	38.08	23.02	63.97	25.69	17.06	32.11	15.69	34.85	41.74	23.77	28.58	31.12
• TENT (ICLR'21)	21.01	29.85	9.47	54.05	12.89	3.92	21.97	0.56	22.93	16.45	16.41	19.95	19.12
• SAR (ICLR'23)	19.41	32.51	10.69	54.81	12.93	3.48	21.55	2.48	23.98	26.95	12.82	16.57	19.85
• DUO (ICCV'25)	29.34	37.45	24.11	62.85	24.83	16.66	31.08	12.51	31.63	38.92	24.06	28.42	30.16
• ReCAP (ICML'25)	29.52	38.52	23.08	62.28	26.00	17.59	32.42	15.73	34.92	40.06	24.25	29.05	31.12
• ZeroSiam (ICLR'26)	30.01	38.41	23.51	63.69	26.09	18.20	32.16	16.11	35.06	42.29	24.34	29.14	31.58
• TRACE (Ours)	**33.45**	**40.12**	**24.71**	**64.78**	**27.62**	**19.31**	**33.05**	**17.98**	**36.10**	**43.23**	**27.25**	**31.67**	**33.27**
PartSTAD (Source)	35.48	43.89	23.54	65.80	30.50	21.90	39.89	18.82	35.42	43.23	26.10	31.41	34.67
• TENT (ICLR'21)	23.52	40.33	13.61	55.02	27.02	9.80	30.38	6.24	26.89	16.65	25.21	12.60	23.94
• SAR (ICLR'23)	28.21	38.69	9.89	59.97	26.20	14.95	31.72	11.89	29.72	24.65	23.68	19.75	26.61
• DUO (ICCV'25)	34.05	41.58	20.58	65.06	30.83	21.52	38.78	17.91	36.09	40.86	27.95	32.11	33.94
• ReCAP (ICML'25)	34.58	43.75	23.24	66.70	32.23	22.15	40.08	19.05	37.46	43.91	26.25	30.89	35.02
• ZeroSiam (ICLR'26)	35.52	44.10	23.66	67.90	31.69	22.03	40.15	18.96	37.84	42.27	28.85	32.20	35.43
• TRACE (Ours)	**37.32**	**46.85**	**23.98**	**71.44**	**35.66**	**26.30**	**43.53**	**21.85**	**40.04**	**44.66**	**33.85**	**36.61**	**38.51**

5.2 Evaluation Results

Evaluation on Corruption Shifts. We evaluate robustness against six point-cloud corruptions at severity levels 3 and 5, as shown in Tab. 1. 1) *Severe Degradation under Corruptions.* Increasing severity from level 3 to 5 reduces source-model mIoU by 7.1% on average, highlighting the need for test-time adaptation. 2) *Stable Gains under Moderate Corruption.* At level 3, TRACE consistently outperforms all baselines across corruption types, improving mIoU by **+2.8%** on both base models. 3) *Robustness under Severe Corruption.* At level 5, existing methods often underperform the frozen models or yield marginal gains ($<0.4\%$ mIoU). In contrast, TRACE remains effective and achieves about **7**$\times$ and **10**$\times$ larger gains than the latest SOTA on the two base models, demonstrating its robustness under severely degraded target geometry.

Evaluation on Category Shifts. We evaluate six bidirectional category-shift tasks, with results reported in Tab. 2. 1) Existing methods suffer from negative transfer on the Composition and Shape tasks, performing worse than the source model. This indicates that category shifts introduce unreliable cross-view evidence that can mislead conventional objectives. 2) TRACE consistently improves performance across all transfer directions. Compared with prior methods, it increases average mIoU by **+2.7%** on PartSLIP and **+3.5%** on PartSTAD, showing stable adaptation under category shifts.

Evaluation on Sim-Real Shifts. We evaluate sim-to-real adaptation on ScanObjectNN, as reported in Tab. 3. Prior methods can degrade severely on real scans; for instance, TENT and SAR collapse to near-zero on the Shelf and Display. In contrast, TRACE consistently yields positive gains across all tasks, improving mIoU by **+1.7%** and **+3.1%** on the two base models. These gains indicate that TRACE can recover reliable adaptation signals from noisy and incomplete real-world geometry.

Running Time Comparison. For a fair efficiency comparison, we cache the rendered multi-view images and measure only the runtime from multi-view inputs to final 3D predictions. As shown in Tab. 2, our TRACE achieves significantly superior performance with the second-best time cost (just +1s compared to ZeroSiam), highlighting its efficiency.

6 Ablation Study and Visualization

In this section, without loss of generality, we conduct ablation studies and visualizations under category shifts using PartSTAD as the base model. Focusing on pivotal components of TRACE, we perform various experiments to analyze their impacts. Refer to Appendix C for more experiments.

6.1 Sensitivity Analysis

We assess the robustness of TRACE to three key configurations: the uncertainty budget m, the gate threshold τ, and the regularization weight λ.

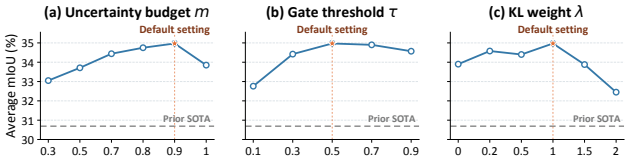

Figure 4: Performance under varying hyperparameters: (a) uncertainty mass m, (b) threshold τ, and (c) KL weight λ.

As shown in Fig. 4(a), increasing the uncertainty budget m improves performance by enhancing TRACE's sensitivity to ambiguous evidence, while overly large values (*e.g.*, $m > 1.0$) suppress informative semantic cues and yield diminishing gains. Fig. 4(b) shows stable performance across a wide range of gating thresholds τ, indicating that TRACE is insensitive to the precise gating criterion. Similarly, Fig. 4(c) shows that moderate λ values achieve the best balance, whereas overly weak or strong regularization slightly degrades performance. Overall, TRACE maintains consistent gains across all three hyperparameters, demonstrating strong robustness to hyperparameter choices.

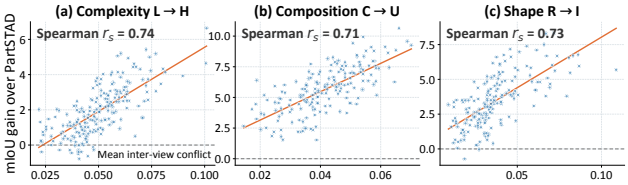

Figure 5: Relationship between inter-view conflict and performance gain by TRACE on category shift benchmarks.

6.2 Impact of Multi-View Evidence Conflict

We further examine how TRACE behaves as multi-view evidence becomes increasingly conflicting. For each instance, we compute the average semantic disagreement across valid view pairs, where higher scores indicate stronger conflict. As shown in Fig. 5, TRACE's instance-level adaptation gain increases consistently with conflict, with Spearman correlations above 0.7 across all three settings. Low-conflict instances are often already well resolved by cross-view agreement, whereas TRACE provides larger gains when views disagree. This trend supports our motivation that structured evidence reconciliation is particularly beneficial when multi-view predictions are conflicting.

6.3 Real-World Demo

To validate the real-world applicability, we evaluate TRACE with a LiDAR-equipped iPhone. As shown in Fig. 6, it exhibits consistent gains with our benchmark results. In the first row, inconsistent assignments near part boundaries lead to conflicting predictions that prior methods fail to reconcile, whereas TRACE leverages cross-view consensus to recover sharper boundaries. In the second row, ambiguous predictions over large regions are amplified in ZeroSiam; TRACE instead suppresses uncertain evidence and selectively reinforces reliable hypotheses, recovering correct part labels.

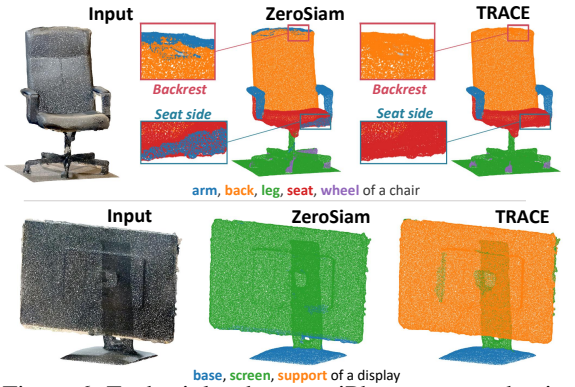

Figure 6: Each triplet shows an iPhone-captured point cloud, segmentation results of ZeroSiam and TRACE.

7 Conclusion

In this work, we study test-time adaptation for 2D-to-3D part segmentation under distribution shifts and establish the first benchmark covering corruption, category, and sim-to-real scenarios. To address the evidence-coupling dilemma, we propose TRACE, which reformulates adaptation as tracing and reinforcing compatible evidence chains across multi-view predictions. Extensive experiments across diverse distribution shifts, backbone models, and real-world LiDAR scans consistently demonstrate the robustness and practical effectiveness of TRACE. We hope this work inspires future research on consensus-aware adaptation, extending this insight to broader multi-view domains.

AI USE STATEMENT

We use LLM as an assistant to help us polish our writing. The tool was not used to generate experimental data or reported results. All AI-assisted revisions were reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of this work. We are grateful for the convenience that LLMs has brought to research writing and call for the rational utilization.

ETHICS STATEMENT

This work uses publicly available datasets of 3D objects and their annotations in accordance with the applicable licenses and terms of use. It does not involve human-subject research or the collection or processing of personal, confidential, or otherwise sensitive data. The evaluated models neither use demographic attributes nor make decisions about individuals; therefore, we identify no direct privacy or human-subject risks arising from the experimental protocol. The authors declare no financial, professional, or personal conflicts of interest related to this work. We have reviewed and followed the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of the experimental results, We provide the information needed to reproduce our method and experiments throughout the main paper and appendix. Specifically, Appendix D.4 presents pseudocode for the TRACE inference and adaptation procedure; Sec. 5.1 reports the hyperparameter configurations; Appendix B details the benchmark construction, source–target splits, and evaluation protocols; and Appendix D describes the base models and implementation settings. Complete proofs of the theoretical results are provided in Appendix A.

REFERENCES

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

Guohao Chen, Shuaicheng Niu, Deyu Chen, Jiahao Yang, Zitian Zhang, Mingkui Tan, Pengcheng Wu, and Zhiqi Shen. Zerosiam: An efficient asymmetry for test-time entropy optimization without collapse. In *Third Workshop on Test-Time Updates (Main Track)*, 2026.

Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5881–5892. IEEE, 2025.

Ken Deng, Yunhan Yang, Jingxiang Sun, Xihui Liu, Yebin Liu, Ding Liang, and Yan-Pei Cao. Geosam2: Unleashing the power of sam2 for 3d part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6367–6376, 2026.

Nikita Durasov, Assaf Shocher, Doruk Oner, Gal Chechik, Alexei A Efros, and Pascal Fua. IT^3: Idempotent test-time training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 14867–14883. PMLR, 13–19 Jul 2025.

Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7081–7091. IEEE, 2023.

Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with

only 2d image supervision. In *2021 International Conference on 3D Vision (3DV)*, pp. 361–372. IEEE, 2021.

Gustavo A Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Ismail Ben Ayed, and Christian Desrosiers. Clust3: Information invariant test-time training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6136–6145, 2023.

Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Ling-Yu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. *arXiv preprint arXiv:2505.20704*, 2025a.

Zixuan Hu, Dongxiao Li, Xinzhu Ma, Shixiang Tang, Xiaotong Li, Wenhan Yang, and Ling-Yu Duan. Adaptive dual uncertainty optimization: Boosting monocular 3d object detection under test-time shifts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7273–7283. IEEE, 2025b.

Zixuan Hu, Xuantuo Huang, Yancheng Li, Yichun Hu, Shengyong Xu, and Ling-Yu Duan. Turning adaptation into assets: Cross-domain bridging for online vision-language navigation. *arXiv preprint arXiv:2605.23257*, 2026.

Hyunjin Kim and Minhyuk Sung. Partstad: 2d-to-3d part segmentation task adaptation. In *European conference on computer vision*, pp. 422–439. Springer, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4558–4567. IEEE, 2018.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023.

Hongbin Lin, Yifan Zhang, Shuaicheng Niu, Shuguang Cui, and Zhen Li. Monotta: Fully test-time adaptation for monocular 3d object detection. In *European Conference on Computer Vision*, pp. 96–114. Springer, 2024.

Jian Liu, Xiaoshui Huang, Tianyu Huang, Lu Chen, Yuenan Hou, Shixiang Tang, Ziwei Liu, Wanli Ouyang, Wangmeng Zuo, Junjun Jiang, et al. A comprehensive survey on 3d content generation. *arXiv preprint arXiv:2402.01166*, 2024a.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21736–21746. IEEE, 2023.

Weihuang Liu, Xi Shen, Haolun Li, Xiuli Bi, Bo Liu, Chi-Man Pun, and Xiaodong Cun. Depth-aware test-time training for zero-shot video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19218–19227, 2024b.

Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-sam: Native 3d part segmentation. *arXiv preprint arXiv:2509.06784*, 2025.

Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019.

John W Moon and Leo Moser. On cliques in graphs. *Israel journal of Mathematics*, 3(1):23–28, 1965.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International Conference on Machine Learning (ICML)*, 2022.

Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16928–16937, 2022.

Haopeng Sun, Lumin Xu, Sheng Jin, Ping Luo, Chen Qian, and Wentao Liu. Program: Prototype graph model based pseudo-label learning for test-time adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 34298–34328, 2024.

Yun-Yun Tsai, Fu-Chen Chen, Albert YC Chen, Junfeng Yang, Che-Chun Su, Min Sun, and Cheng-Hao Kuo. Gda: Generalized diffusion for robust test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23242–23251, 2024.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1588–1597. IEEE, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2569–2578. IEEE, 2018.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in neural information processing systems*, 35:33330–33342, 2022.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11094–11104. IEEE, 2020.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6757–6767. IEEE, 2023.

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

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

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.

Zhi Zhou, Kun-Yang Yu, Lan-Zhe Guo, and Yu-Feng Li. Fully test-time adaptation for tabular data. In *Proceedings of the 39th AAAI conference on Artificial Intelligence*, 2025.

Tracing Chains of Evidence: Cross-View Test-Time Adaptation for Point Cloud Part Segmentation

————Appendix————

The structure of Appendix is as follows:

- Appendix A contains all proofs and propositions for theoretical claims in the main manuscript.
- Appendix B details the benchmark design and data construction.
- Appendix C provides additional experimental results on TTA strategies and ablation studies on component effectiveness.
- Appendix D outlines further implementation details of our method and baselines.

A Theoretical Analysis and Proofs

Below, we will provide detailed proofs of the theoretical results in the main manuscript.

Notation. We first recall the notation used in the main manuscript and this appendix. For a fixed 3D point $\mathbf{p}$, let $\mathcal{O}$ be its nonempty set of valid views and let $\mathcal{S} = \{1, \dots, C, \perp\}$ be the state space, where $\perp$ denotes uncertainty. For each view $v \in \mathcal{O}$ and state $s \in \mathcal{S}$, $\bar{w}_v(s) = \bar{w}(v, s)$ denotes the calibrated vertex weight defined in Eq. 8.

For any nonempty view subset $\mathcal{A} \subseteq \mathcal{O}$, let $\mathcal{G}[\mathcal{A}]$ be the subgraph induced by the vertices associated with the views in $\mathcal{A}$, and let $\mathcal{K}(\mathcal{A})$ be the set of its maximum cliques. Each $K \in \mathcal{K}(\mathcal{A})$ selects one state from every view in $\mathcal{A}$ and induces a hypothesis $\ell(K) \in \mathcal{S}$. We define

$$\mathcal{K}_s(\mathcal{A}) = \{K \in \mathcal{K}(\mathcal{A}) : \ell(K) = s\}, \qquad Z_s^{\mathcal{A}} = \sum_{K \in \mathcal{K}_s(\mathcal{A})} \prod_{(v,t) \in K} \bar{w}_v(t), \tag{14}$$

where $Z_s^{\mathcal{A}}$ is the unnormalized support assigned to hypothesis s by the views in $\mathcal{A}$. Its normalized counterpart is $\widehat{Z}_s^{\mathcal{A}} = Z_s^{\mathcal{A}} / D^{\mathcal{A}}$, with $D^{\mathcal{A}} = \sum_{t \in \mathcal{S}} Z_t^{\mathcal{A}}$. For the empty view set, we adopt the neutral convention $Z_{\perp}^{\varnothing} = 1$ and $Z_c^{\varnothing} = 0$ for every semantic class c. We omit the superscript when $\mathcal{A} = \mathcal{O}$.

A.1 Proof of Proposition 1

Proposition 1 (Binary Composition Rule). *Let $\mathcal{A}$ and $\mathcal{B}$ be two disjoint subsets of valid views, and let $Z^{\mathcal{A}}$ and $Z^{\mathcal{B}}$ be the hypothesis-level supports obtained from their induced subgraphs. For every semantic class $c \in \{1, \dots, C\}$, the support over $\mathcal{A} \cup \mathcal{B}$ satisfies*

$$Z_{\perp}^{\mathcal{A}\cup\mathcal{B}} = Z_{\perp}^{\mathcal{A}} Z_{\perp}^{\mathcal{B}}, \qquad Z_c^{\mathcal{A}\cup\mathcal{B}} = Z_c^{\mathcal{A}} Z_{\perp}^{\mathcal{B}} + Z_{\perp}^{\mathcal{A}} Z_c^{\mathcal{B}} + Z_c^{\mathcal{A}} Z_c^{\mathcal{B}}. \tag{15}$$

The composition is commutative and associative.

Proof. We prove the result by characterizing the maximum cliques of the compatibility graph and partitioning them according to their restrictions to $\mathcal{A}$ and $\mathcal{B}$.

If $\mathcal{A} = \varnothing$ or $\mathcal{B} = \varnothing$, Eq. 15 follows immediately from the neutral convention above. We therefore consider the nonempty case in the remainder of the proof.

Structure of a Complete Evidence Chain. Vertices associated with the same view are disconnected, so a clique contains at most one vertex from each view. The uncertainty vertices from different views are pairwise adjacent, which constructs a clique containing one vertex from every view. Every maximum clique of $\mathcal{G}[\mathcal{A}]$ has size $|\mathcal{A}|$ and selects exactly one state from each view.

The cross-view edge rule further constrains the selected states. Two semantic vertices are adjacent only when they indicate the same class, whereas an uncertainty vertex is compatible with both semantic and uncertainty vertices from other views. Hence, a maximum clique either selects $\perp$ in every view or contains a unique semantic class c and selects one state from $\{c, \perp\}$ in every view. The former induces the uncertainty hypothesis, while the latter induces the class-c hypothesis.

Decomposition over Two Disjoint View Sets. For any clique $K \in \mathcal{K}(\mathcal{A} \cup \mathcal{B})$, let $K_{\mathcal{A}}$ and $K_{\mathcal{B}}$ be its restrictions to the vertices from $\mathcal{A}$ and $\mathcal{B}$, respectively. Since $\mathcal{A}$ and $\mathcal{B}$ are disjoint and K contains one vertex from every view, $K_{\mathcal{A}} \in \mathcal{K}(\mathcal{A})$ and $K_{\mathcal{B}} \in \mathcal{K}(\mathcal{B})$. Conversely, two restricted cliques form a clique over $\mathcal{A} \cup \mathcal{B}$ exactly when their induced hypotheses are compatible.

For the uncertainty hypothesis, both restrictions must be all-uncertainty cliques, giving the bijection

$$\mathcal{K}_{\perp}(\mathcal{A} \cup \mathcal{B}) \cong \mathcal{K}_{\perp}(\mathcal{A}) \times \mathcal{K}_{\perp}(\mathcal{B}). \tag{16}$$

For a semantic class c, the two restrictions can induce $(c, \perp)$, $(\perp, c)$, or (c, c). These three cases are mutually exclusive and exhaust all class-c cliques. Therefore,

$$\begin{aligned}\mathcal{K}_c(\mathcal{A} \cup \mathcal{B}) \cong \big(\mathcal{K}_c(\mathcal{A}) \times \mathcal{K}_{\perp}(\mathcal{B})\big) \,\dot\cup\, \big(\mathcal{K}_{\perp}(\mathcal{A}) \times \mathcal{K}_c(\mathcal{B})\big) \\ \dot\cup\, \big(\mathcal{K}_c(\mathcal{A}) \times \mathcal{K}_c(\mathcal{B})\big),\end{aligned} \tag{17}$$

where $\dot\cup$ denotes a disjoint union.

Factorization of Clique Support. The weight of a clique is the product of its vertex weights. Because $\mathcal{A}$ and $\mathcal{B}$ are disjoint, this product factorizes as

$$\prod_{(v,t)\in K} \bar{w}_v(t) = \left(\prod_{(v,t)\in K_{\mathcal{A}}} \bar{w}_v(t)\right)\left(\prod_{(v,t)\in K_{\mathcal{B}}} \bar{w}_v(t)\right). \tag{18}$$

Summing Eq. 18 over the bijection in Eq. 16 yields

$$Z_{\perp}^{\mathcal{A}\cup\mathcal{B}} = Z_{\perp}^{\mathcal{A}} Z_{\perp}^{\mathcal{B}}. \tag{19}$$

Similarly, summing over the three disjoint cases in Eq. 17 gives

$$Z_c^{\mathcal{A}\cup\mathcal{B}} = Z_c^{\mathcal{A}} Z_{\perp}^{\mathcal{B}} + Z_{\perp}^{\mathcal{A}} Z_c^{\mathcal{B}} + Z_c^{\mathcal{A}} Z_c^{\mathcal{B}}, \tag{20}$$

which proves the binary composition rule.

Commutativity and Associativity. Define the inclusive support for class c as

$$T_c^{\mathcal{A}} := Z_{\perp}^{\mathcal{A}} + Z_c^{\mathcal{A}}. \tag{21}$$

Equations 19 and 20 imply

$$Z_{\perp}^{\mathcal{A}\cup\mathcal{B}} = Z_{\perp}^{\mathcal{A}} Z_{\perp}^{\mathcal{B}}, \qquad T_c^{\mathcal{A}\cup\mathcal{B}} = T_c^{\mathcal{A}} T_c^{\mathcal{B}}. \tag{22}$$

Thus, both $Z_{\perp}$ and T_c compose by scalar multiplication. The composition is commutative and associative in these coordinates, and the original semantic support is recovered by $Z_c^{\mathcal{A}} = T_c^{\mathcal{A}} - Z_{\perp}^{\mathcal{A}}$. The same properties therefore hold for the binary rule in Eq. 15. □

A.2 Proof of Proposition 2

Proposition 2 (Multi-View Closed-Form Composition). *Let $\mathcal{G}$ be the evidence graph for point $\mathbf{p}$ with valid views $\mathcal{O}$. For each $s \in \mathcal{S}$, the normalized maximum-clique support is*

$$\widehat{Z}_s = \begin{cases} \dfrac{\prod_{v\in\mathcal{O}} \bar{w}_v(\perp)}{\sum_{c=1}^{C} \prod_{v\in\mathcal{O}} \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - (C-1)\prod_{v\in\mathcal{O}} \bar{w}_v(\perp)}, & s = \perp, \\ \dfrac{\prod_{v\in\mathcal{O}} \big(\bar{w}_v(s) + \bar{w}_v(\perp)\big) - \prod_{v\in\mathcal{O}} \bar{w}_v(\perp)}{\sum_{c=1}^{C} \prod_{v\in\mathcal{O}} \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - (C-1)\prod_{v\in\mathcal{O}} \bar{w}_v(\perp)}, & s \neq \perp . \end{cases} \tag{23}$$

Proof. We first derive the unnormalized support for an arbitrary nonempty view subset $\mathcal{A} \subseteq \mathcal{O}$ and then normalize the result. We prove by induction on $n = |\mathcal{A}|$ that

$$Z_{\perp}^{\mathcal{A}} = \prod_{v\in\mathcal{A}} \bar{w}_v(\perp), \tag{24}$$

$$Z_c^{\mathcal{A}} = \prod_{v\in\mathcal{A}} \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - \prod_{v\in\mathcal{A}} \bar{w}_v(\perp), \qquad c \in \{1, \ldots, C\}. \tag{25}$$

Base Case. Let $\mathcal{A} = \{v\}$ contain a single view. Its induced graph has the vertices $\{(v, s) : s \in \mathcal{S}\}$ and no edges between them because states from the same view are mutually exclusive. Hence, every vertex forms a maximum clique of size one. For each state $s \in \mathcal{S}$,

$$\mathcal{K}_s(\{v\}) = \big\{\{(v, s)\}\big\}. \tag{26}$$

Substituting Eq. 26 into the support definition in Eq. 14 removes both the sum over cliques and the product over vertices:

$$Z_s^{\{v\}} = \sum_{K \in \mathcal{K}_s(\{v\})} \prod_{(u,t)\in K} \bar{w}_u(t) = \prod_{(u,t)\in\{(v,s)\}} \bar{w}_u(t) = \bar{w}_v(s). \tag{27}$$

Taking $s = \perp$ gives $Z_\perp^{\{v\}} = \bar{w}_v(\perp)$, which is Eq. 24 for one view. For a semantic class c, Eq. 25 reduces to

$$\prod_{u\in\{v\}} \big(\bar{w}_u(c) + \bar{w}_u(\perp)\big) - \prod_{u\in\{v\}} \bar{w}_u(\perp) = \bar{w}_v(c) + \bar{w}_v(\perp) - \bar{w}_v(\perp) = \bar{w}_v(c) = Z_c^{\{v\}}.$$

Thus, both closed-form branches hold for $|\mathcal{A}| = 1$.

Inductive Step. Assume that Eqs. 24 and 25 hold for a subset $\mathcal{A}$ containing n views. Let $v \notin \mathcal{A}$ and define $\mathcal{A}' = \mathcal{A} \cup \{v\}$. The singleton result above gives $Z_\perp^{\{v\}} = \bar{w}_v(\perp)$ and $Z_c^{\{v\}} = \bar{w}_v(c)$.

The union $\mathcal{A}'$ induces the uncertainty hypothesis only when both $\mathcal{A}$ and the added view v select uncertainty. Applying the uncertainty branch of Proposition 1 and then the induction hypothesis gives

$$Z_\perp^{\mathcal{A}'} = Z_\perp^{\mathcal{A}} Z_\perp^{\{v\}} = \left(\prod_{u\in\mathcal{A}} \bar{w}_u(\perp)\right) \bar{w}_v(\perp) = \prod_{u\in\mathcal{A}'} \bar{w}_u(\perp), \tag{28}$$

which establishes Eq. 24 for $n + 1$ views.

For a semantic class c, the union induces class c through exactly three cases: $(c, \perp)$, $(\perp, c)$, and (c, c) for the hypotheses contributed by $\mathcal{A}$ and $\{v\}$, respectively. Proposition 1 therefore gives

$$\begin{aligned} Z_c^{\mathcal{A}'} &= Z_c^{\mathcal{A}} Z_\perp^{\{v\}} + Z_\perp^{\mathcal{A}} Z_c^{\{v\}} + Z_c^{\mathcal{A}} Z_c^{\{v\}} \\ &= Z_c^{\mathcal{A}} \bar{w}_v(\perp) + Z_\perp^{\mathcal{A}} \bar{w}_v(c) + Z_c^{\mathcal{A}} \bar{w}_v(c) \\ &= Z_c^{\mathcal{A}} \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) + Z_\perp^{\mathcal{A}} \bar{w}_v(c) \\ &= \big(Z_c^{\mathcal{A}} + Z_\perp^{\mathcal{A}}\big)\big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - Z_\perp^{\mathcal{A}} \bar{w}_v(\perp) \end{aligned} \tag{29}$$

The last line adds and subtracts $Z_\perp^{\mathcal{A}} \bar{w}_v(\perp)$. This rearrangement isolates the inclusive support $Z_c^{\mathcal{A}} + Z_\perp^{\mathcal{A}}$, whose two subtraction terms cancel under the induction hypothesis:

$$\begin{aligned} Z_c^{\mathcal{A}} + Z_\perp^{\mathcal{A}} &= \left[\prod_{u\in\mathcal{A}} \big(\bar{w}_u(c) + \bar{w}_u(\perp)\big) - \prod_{u\in\mathcal{A}} \bar{w}_u(\perp)\right] + \prod_{u\in\mathcal{A}} \bar{w}_u(\perp) \\ &= \prod_{u\in\mathcal{A}} \big(\bar{w}_u(c) + \bar{w}_u(\perp)\big). \end{aligned} \tag{30}$$

Substituting Eq. 30 and the uncertainty induction hypothesis into Eq. 29 yields

$$\begin{aligned} Z_c^{\mathcal{A}'} &= \left[\prod_{u\in\mathcal{A}} \big(\bar{w}_u(c) + \bar{w}_u(\perp)\big)\right] \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - \left[\prod_{u\in\mathcal{A}} \bar{w}_u(\perp)\right] \bar{w}_v(\perp) \\ &= \prod_{u\in\mathcal{A}'} \big(\bar{w}_u(c) + \bar{w}_u(\perp)\big) - \prod_{u\in\mathcal{A}'} \bar{w}_u(\perp). \end{aligned} \tag{31}$$

This is Eq. 25 for $n + 1$ views. Together with Eq. 28, it completes the induction.

Normalization. Setting $\mathcal{A} = \mathcal{O}$ and summing the unnormalized supports over all states gives

$$D^{\mathcal{O}} = Z_\perp^{\mathcal{O}} + \sum_{c=1}^{C} Z_c^{\mathcal{O}} = \sum_{c=1}^{C} \prod_{v\in\mathcal{O}} \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - (C - 1) \prod_{v\in\mathcal{O}} \bar{w}_v(\perp). \tag{32}$$

The all-uncertainty chain has positive weight because $\bar{w}_v(\perp) > 0$ for every valid view, so $D^{\mathcal{O}} > 0$. Dividing Eqs. 24 and 25 by Eq. 32 gives the two branches of Eq. 23.

□



Figure 7: Visualization of six geometric corruption types in PartNet-C at severity levels 3 (top row) and 5 (bottom row). The original point cloud is shown on the left.

B Benchmark Construction and Evaluation Protocol

This section provides the benchmark construction details and evaluation protocols for the three distribution-shift scenarios considered in the main paper: corruption shift, category shift, and simulation-to-real shift. We first summarize how the source and target domains are defined under each scenario. We then describe the construction of PartNet-C, including its six geometric corruption operators and severity settings; the criteria and category assignments used to form the six directional category-shift tasks; and the preparation of the ScanObjectNN target domains, including the category-specific label spaces, text prompts, and separation between source training and target-time adaptation. Together, these details specify how each shift is instantiated while ensuring that target-domain annotations are reserved exclusively for evaluation.

B.1 Overview of Three Scenarios

1) **Corruption Shift.** Following PointCloud-C (Ren et al., 2022), we construct PartNet-C, a corrupted version of PartNet-Mobility for evaluating robustness under geometric perturbations. To emulate variations in point-cloud acquisition quality, we apply six corruption operators, including Local and Global Dropout, Local and Global Addition, Rotation, and Jitter, each with five severity levels; higher levels correspond to stronger corruption shifts. The original training split is used as the source domain, while the corrupted test split serves as the target domain. This setting keeps the semantic space fixed and isolates the impact of geometric degradation, providing a controlled testbed for evaluating whether test-time adaptation can recover reliable part predictions from corrupted target geometry.

2) **Category Shift.** This setting is built on PartNet-Mobility (Xiang et al., 2020), a part segmentation dataset containing object point clouds with point-wise semantic annotations. It evaluates cross-category generalization, where source and target objects differ in structural and semantic properties. We organize the categories according to three attributes: object complexity, part-composition commonality, and shape regularity. For each attribute, we form two disjoint category groups with 15 and 10 categories, respectively, and evaluate both transfer directions, yielding six tasks: **L**ow $\leftrightarrow$ **H**igh complexity, **C**ommon $\leftrightarrow$ **U**ncommon composition, and **R**egular $\leftrightarrow$ **I**rregular shape.

3) **Sim-Real Shift.** To evaluate generalization to real-world observations, we introduce a Sim-Real shift setting based on ScanObjectNN (Uy et al., 2019), a real-world point-cloud dataset. Unlike simulation dataset which consists of clean synthetic CAD models, ScanObjectNN contains 2,902 objects extracted from indoor scans and exhibits realistic geometric variations caused by scanning noise, incompleteness, and occlusion. We use the model trained on PartNet-Mobility as the source model and treat each ScanObjectNN category with more than one semantic part as an independent target domain, resulting in 12 adaptation tasks. This setting defines a challenging and practical adaptation scenario with mixed shifts, including visual-domain discrepancies between synthetic and real observations as well as geometric degradation from real-world scanning.

B.2 CORRUPTION SHIFT CONSTRUCTION

We construct PartNet-C from the test split of PartNet-Mobility by applying six geometric corruption operators that represent common degradations in point-cloud acquisition. These corruptions cover four major sources of distribution shift: spurious observations, missing geometry, changes in object pose, and point-level localization noise. The semantic labels of the retained points remain unchanged, while newly injected perturbation points are treated as unlabeled observations. The six corruption operators are defined below.

Global and Local Addition. The addition corruptions introduce spurious points that do not belong to the underlying object geometry. Global Addition uniformly samples unlabeled perturbation points throughout the spatial volume enclosing the entire point cloud, adding 30 and 50 points per point cloud at severity levels 3 and 5, respectively. These scattered outliers alter the global point-density distribution and may produce isolated artifacts in the rendered views, thereby testing whether the model can distinguish genuine object geometry from spatially dispersed noise. Local Addition instead inserts a total of 300 unlabeled points per point cloud at severity level 3, with the count increasing to 500 at level 5. The total addition count is randomly allocated across one to seven local clusters. For each cluster, we randomly select an original point as its center and generate new points by adding Gaussian offsets. Because these points are spatially concentrated, they can form dense artificial structures near genuine object parts and interfere with local boundaries or projected silhouettes. This corruption therefore models localized interference, such as reconstruction artifacts concentrated around a small region.

Global and Local Dropout. The dropout corruptions remove observed points while preserving the coordinates and semantic labels of the remaining points. Global Dropout randomly permutes the input points and retains the first $\lfloor N(1-d) \rfloor$ points, where N is the original number of points and d is the dropout rate. The dropout rate is set to $d = 0.50$ at severity level 3 and $d = 0.75$ at level 5. The resulting sparsity can weaken part boundaries and reduce the amount of geometric evidence available across rendered views. Local Dropout removes a total of 300 and 500 points per point cloud at severity levels 3 and 5, respectively. The total removal count is randomly allocated across one to seven local clusters. For each cluster, an anchor point is randomly selected, and the allocated number of points closest to the anchor is removed, producing contiguous holes or partially missing surfaces. Unlike Global Dropout, which affects the object broadly, Local Dropout can completely suppress the observation of a small part or a discriminative local structure. It therefore models localized sensing failures, partial occlusion, and incomplete reconstruction.

Rotation. Rotation applies a rigid rotation to the complete point cloud. For each object, the rotation angles about the three coordinate axes are independently sampled from a uniform distribution over $[-18°, 18°]$ at severity level 3 or $[-30°, 30°]$ at severity level 5. The three axis rotations are then composed to obtain the transformation applied to all points. This operation changes the global orientation of the object while preserving its internal geometry, point correspondences, and semantic labels. Because the rendering cameras remain fixed, rotating the object changes which surfaces are visible and where individual parts appear in each image. Rotation therefore evaluates sensitivity to observation pose and tests whether semantic evidence can be consistently associated across views when the target object is presented in an orientation different from those observed in the source domain.

Jitter. Jitter independently adds zero-mean Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to each coordinate of each point, displacing it within a local neighborhood of its original position. The noise standard deviation increases from 0.03 at severity level 3 to 0.05 at level 5. Unlike dropout, jitter preserves the number of observed points, but perturbs their local spatial arrangement. The resulting deformation can reduce surface smoothness, blur fine-grained geometric boundaries, and introduce irregularities into rendered contours. This corruption models point-level localization errors caused by sensor noise, depth estimation errors, or imperfect 3D reconstruction.

For each corruption operator, we define five severity levels that progressively increase the corresponding perturbation strength. For Global and Local Addition, severity controls the number of injected points or local perturbation clusters. For Global and Local Dropout, it controls the number of removed points or the spatial extent of the missing regions. For Rotation and Jitter, severity controls the rotation magnitude and coordinate-perturbation magnitude, respectively. We evaluate severity levels 3 and 5, representing moderate and severe geometric degradation. As shown in Fig. 7,

Table 4: Category assignments for the category-shift benchmark, organized by object complexity, composition commonality, and shape regularity. Numbers in parentheses indicate the number of categories in each group.

Complexity	Composition	Shape
Low (15)	**Common** (15)	**Regular** (15)
Bottle, Box, Bucket, Camera, Cart, Clock, Dishwasher, Eyeglasses, FoldingChair, Pen, Pliers, Remote, Suitcase, Switch, Window	Box, Bucket, Dishwasher, Dispenser, Door, Faucet, Kettle, KitchenPot, Knife, Oven, Refrigerator, Scissors, Stapler, Switch, TrashCan	Box, Bucket, Dishwasher, Display, Door, Globe, Kettle, Keyboard, Microwave, Oven, Printer, Refrigerator, Toaster, TrashCan, Window
High (10)	**Uncommon** (10)	**Irregular** (10)
Chair, Display, Lamp, Laptop, Microwave, Mouse, Safe, Scissors, Toilet, TrashCan	Camera, Clock, Display, Keyboard, Laptop, Mouse, Pen, Pliers, USB, Window	Cart, Dispenser, FoldingChair, Lamp, Mouse, Pen, Pliers, Scissors, Stapler, Switch

these two settings allow us to assess both performance degradation under realistic perturbations and robustness under more challenging acquisition conditions.

B.3 Category Shift Construction

We construct category-shift tasks along three complementary dimensions: object complexity, part-composition commonality, and shape regularity. Each dimension partitions the object categories into two disjoint groups and defines two transfer directions between them. The resulting category assignments are reported in Tab. 4. The three grouping criteria are detailed below.

Object Complexity. We define object complexity by the number of semantic part types an object contains. Object categories with at most two part types are assigned to the **L**ow-complexity group, while those with at least three part types are assigned to the **H**igh-complexity group. This grouping evaluates transfer between objects with different numbers of semantic part types.

Part-Composition Commonality. We count the occurrences of semantic part names across the entire dataset to distinguish common from uncommon part compositions. Part names occurring at least three times are considered common, while the remaining part names are considered uncommon. Object categories containing common parts are assigned to the **C**ommon-composition group, and all other categories are assigned to the **U**ncommon-composition group. This grouping evaluates generalization between common and uncommon parts.

Shape Regularity. We divide the object categories into **R**egular- and **I**rregular-shape groups through a category-level qualitative assessment of their geometry. The assessment considers the consistency of the global form and part arrangement across instances, including whether a category exhibits a stable canonical layout and approximate symmetry. Categories with relatively consistent and structured geometry are assigned to the Regular-shape group. Categories exhibiting stronger asymmetry, articulation, structural variability, or non-canonical part arrangements are assigned to the Irregular-shape group. It evaluates whether cross-category adaptation remains effective when the geometric organization of the target objects differs from that of the source categories.

B.4 Sim-to-Real Shift Construction

ScanObjectNN provides object-level category annotations and, in its raw part-annotated release, point-wise semantic labels for object parts. The point-wise part annotations are used exclusively as ground truth for part segmentation evaluation and are never exposed during test-time adaptation. For each target object category, we construct the candidate label set according to ScanObjectNN's native semantic part taxonomy and incorporate the corresponding part names into the text prompts. All source-side learnable components are trained on synthetic shapes from PartNet-Mobility, while the original 2D-to-3D lifting procedure is kept fixed during adaptation. Consequently, no ScanObjectNN shape or part annotation is used for source training, and the target-domain labels remain

Table 5: Comparison of *view-wise* (VW) and *aggregation-based* (AG) TTA on six category-shift adaptation tasks from PartNet-Mobility, measured by mIoU (%). L/H, C/U, and R/I denote low/high complexity, common/uncommon composition, and regular/irregular shape, respectively.

Model+Method	Complexity			Composition			Shape			**Avg.**
	L→H	H→L	Avg.	C→U	U→C	Avg.	R→I	I→R	Avg.	
PartSLIP (Source)	25.91	20.34	23.13	29.04	15.91	22.48	30.74	16.42	23.58	23.06
• TENT-AG	20.02	20.19	20.10	17.92	6.31	12.12	20.32	9.42	14.86	15.69
• TENT-VW	20.39	19.98	20.19	17.26	6.34	11.80	20.52	9.68	15.10	15.70
• SAR-AG	26.39	20.15	23.27	25.32	4.22	14.76	22.75	8.92	15.84	17.96
• SAR-VW	27.05	20.08	23.57	25.41	4.66	15.04	21.87	8.45	15.16	17.92
• DUO-AG	20.65	20.15	20.40	20.65	9.97	15.31	32.65	15.08	23.87	19.86
• DUO-VW	21.32	19.08	20.20	18.94	9.10	14.02	30.26	15.23	22.75	18.99
• ReCAP-AG	27.69	20.50	24.10	28.80	15.39	22.10	30.12	16.48	23.30	23.17
• ReCAP-VW	27.01	20.38	23.70	27.95	14.68	21.32	29.75	16.62	23.19	22.73
• ZeroSiam-AG	27.29	19.89	23.59	29.74	16.48	23.11	24.01	17.22	20.62	22.44
• ZeroSiam-VW	26.98	20.24	23.61	29.30	16.27	22.79	23.67	17.01	20.34	22.25
• TRACE (Ours)	**30.28**	**21.54**	**25.91**	**30.12**	**17.98**	**24.05**	**37.49**	**17.56**	**27.53**	**25.83**
PartSTAD (Source)	31.71	26.59	29.15	31.92	29.19	30.56	37.95	26.80	32.38	30.69
• TENT-AG	25.14	22.30	23.72	24.46	24.30	24.38	20.39	25.77	23.08	23.73
• TENT-VW	25.31	22.10	23.71	23.86	24.12	23.99	20.10	24.37	22.24	23.31
• SAR-AG	18.14	24.51	21.33	32.61	27.86	30.24	22.76	19.16	20.96	24.18
• SAR-VW	19.01	23.12	21.07	30.98	25.61	28.30	22.84	19.46	21.15	23.51
• DUO-AG	25.04	21.64	23.34	27.55	26.85	27.20	34.53	26.25	30.39	26.97
• DUO-VW	24.74	20.68	22.71	27.22	27.15	27.19	34.00	26.51	30.26	26.72
• ReCAP-AG	33.68	26.05	29.87	32.90	29.22	31.06	37.09	27.05	32.07	31.00
• ReCAP-VW	33.72	25.70	29.71	31.07	29.19	30.13	36.89	27.26	32.08	30.64
• ZeroSiam-AG	35.50	26.45	30.98	29.42	31.35	30.39	38.00	27.91	32.96	31.44
• ZeroSiam-VW	35.01	24.92	29.97	30.19	30.35	30.27	37.59	27.68	32.64	30.96
• TRACE (Ours)	**39.80**	**29.03**	**34.42**	**37.55**	**32.70**	**35.13**	**41.39**	**29.34**	**35.37**	**34.97**

unavailable throughout adaptation. This protocol introduces a realistic simulation-to-real shift from clean synthetic CAD geometry to real-world scans affected by sensor noise, incomplete observations, nonuniform sampling, and occlusion.

C Further Experiments

C.1 View-Wise vs. Aggregation-Based TTA

Existing TTA methods are primarily designed for single-view predictions, leaving open where their adaptation objectives should be applied in a multi-view 2D-to-3D pipeline. We therefore implement each baseline using two strategies that differ in whether adaptation is performed before or after multi-view aggregation. In the *view-wise* (VW) strategy, the TTA objective is computed independently from the prediction of each rendered view before multi-view fusion. Each view therefore contributes a separate adaptation signal. In the *aggregation-based* (AG) strategy, the view-specific predictions are first lifted and fused using the original multi-view aggregation pipeline of the base model. The TTA objective is then computed from the resulting 3D prediction. This strategy relies directly on the output of the original fusion rule.

We evaluate five existing TTA methods under both strategies on the six category-shift tasks. The results in Tab. 5 lead to the following observations:

1) Aggregation Generally Provides a Stronger Adaptation Signal. Across the five compared methods and two base models, AG outperforms its corresponding VW variant in nine of the ten paired comparisons. Averaged over these comparisons, AG improves performance by 0.37 mIoU points. Although the average gain is moderate, its consistency indicates that computing the adaptation objective after multi-view fusion generally provides more informative supervision than treating each rendered view independently. We therefore use AG as the default implementation of these baselines in the main paper.

Table 6: Impact of component Variations. Src., Gra., Geo., and $\mathcal{L}$ denote the frozen source model, compatibility graph, geometric calibration, and the gate of the adaptation objective.

Src.	Gra.	Geo.	$\mathcal{L}$	PartSLIP Cpx.	PartSLIP Cpo.	PartSLIP Shape	PartSLIP Avg.	PartSTAD Cpx.	PartSTAD Cpo.	PartSTAD Shape	PartSTAD Avg.
✔				23.13	22.48	23.58	23.06	29.15	30.56	32.38	30.69
	✔			24.82	23.62	25.80	24.75	31.24	33.62	33.54	32.80
	✔	✔		24.88	<u>23.90</u>	26.21	25.00	32.88	33.97	33.16	33.34
	✔		✔	<u>25.52</u>	23.86	<u>27.09</u>	<u>25.49</u>	<u>34.05</u>	<u>34.62</u>	<u>34.97</u>	<u>34.55</u>
	✔	✔	✔	**25.91**	**24.05**	**27.53**	**25.83**	**34.42**	**35.13**	**35.37**	**34.97**

2) Naive Aggregation Does Not Prevent Negative Transfer. The advantage of AG over VW does not necessarily translate into an improvement over the unadapted source model. On PartSLIP, the AG variants of TENT, SAR, and DUO achieve average mIoUs of 15.69%, 17.96%, and 19.86%, respectively, all below the source performance of 23.06%. The same pattern appears on PartSTAD, where their corresponding averages of 23.73%, 24.17%, and 26.97% remain below the source result of 30.69%. These results show that changing the location of the adaptation objective cannot by itself resolve unreliable cross-view evidence. Direct aggregation may combine correct observations with ambiguous or contradictory cues, causing the resulting errors to be reinforced during adaptation.

3) *TRACE Consistently Outperforms Both Baseline Variants.* TRACE achieves the highest mIoU in every transfer direction under both base models. Compared with the strongest VW baselines, namely ReCAP-VW on PartSLIP and ZeroSiam-VW on PartSTAD, TRACE improves the overall average by **+3.10%** and **+4.01%**, respectively. It also surpasses the strongest AG baselines by +2.66% on PartSLIP and +3.53% on PartSTAD. Moreover, TRACE improves upon the corresponding source models by **+2.77%** and **+4.28%**, whereas several conventional TTA methods exhibit negative transfer. Unlike direct multi-view fusion, TRACE constructs complete evidence chains from semantically compatible hypotheses and calibrates their contributions using geometric consensus. The consistent improvements over both VW and AG variants indicate that effective adaptation requires not only aggregating multiple views, but also explicitly organizing and weighting the evidence according to its cross-view reliability.

C.2 Component Effectiveness

We investigate the impact of individual components by comparing the full method with partial variants, as shown in Tab. 6. Notably, the The results provide three key observations. 1) Geometry-aware calibration consistently improves performance, demonstrating that semantic compatibility alone is insufficient and that grounding cross-view evidence with geometric consensus effectively suppresses unreliable predictions. 2) Adding the gate to the adaptation objective increases the average mIoU on both PartSLIP and PartSTAD. When geometric calibration is enabled, the gate further improves the average mIoU from 25.00% to 25.83% and from 33.34% to 34.97%. These improvements hold across all three category-shift dimensions and both base models. This result verifies the importance of selectively activating the adaptation objective only when the consolidated evidence is sufficiently reliable, thereby reducing the influence of uncertain hypotheses on parameter updates. 3) Combining all three components achieves the best results across adaptation tasks and base models, demonstrating their complementary roles.

D Implementation Details

D.1 Base Models

In the main paper, we evaluate TRACE using two representative 2D-to-3D part segmentation models, PartSLIP (Liu et al., 2023) and PartSTAD (Kim & Sung, 2024). Both models localize text-specified parts in multi-view renderings and lift the detected 2D regions to the input point cloud. TRACE organizes the resulting view-specific semantic and geometric evidence for test-time adaptation. During adaptation, only the learnable prompts are updated, while the pretrained vision-language model and the original 2D-to-3D lifting pipeline remain frozen.

PartSLIP. PartSLIP uses GLIP (Li et al., 2022) as its open-vocabulary 2D detector. Given a rendered view and a category-specific prompt containing the part names, GLIP returns a set of part-labeled

bounding boxes. PartSLIP then converts these boxes into a 3D prediction through a learning-free voting procedure. The input point cloud $\mathcal{P}$ is first over-segmented into superpoints (Landrieu & Simonovsky, 2018) $\{\mathcal{P}_i\}$ using point color and surface-normal features. Let $\nu_v(\mathbf{p}) \in \{0,1\}$ indicate whether point $\mathbf{p}$ is visible in view v, and let $\iota_b(\mathbf{p}) \in \{0,1\}$ indicate whether the projection of $\mathbf{p}$ lies inside box b. If $\mathcal{B}_{v,c}$ denotes the boxes assigned to part class c in view v, PartSLIP computes the class-c support of superpoint $\mathcal{P}_i$ as

$$s_{i,c}^{\mathrm{PS}} = \frac{\sum_{v=1}^{V} \sum_{\mathbf{p}\in\mathcal{P}_i} \nu_v(\mathbf{p}) \mathbb{I}[\exists b \in \mathcal{B}_{v,c} : \iota_b(\mathbf{p}) = 1]}{\sum_{v=1}^{V} \sum_{\mathbf{p}\in\mathcal{P}_i} \nu_v(\mathbf{p})}. \tag{33}$$

The visibility term prevents an occluded point from voting against a box that cannot observe it. Each superpoint receives the class with the largest support. Thus, the superpoint partition supplies a geometric boundary prior, while agreement across projected boxes supplies the semantic evidence used for lifting. The PartSLIP paper additionally studies training-free aggregation of early GLIP visual features through point-level correspondences between views. The released PartSLIP pipeline, which is also used as the starting point of PartSTAD, evaluates the views independently and performs multi-view interaction during 3D voting. We follow this released configuration for the base model.

Within TTA, every PartSLIP bounding box is treated as a region r_i^v. Its GLIP phrase-grounding scores form the semantic vector $\mathbf{s}_i^v$, and the visible 3D points whose projections fall inside the box form its lifted support $\mathcal{Q}_i^v$. These quantities are passed to the cross-view evidence graph before the frozen PartSLIP voting module produces the final 3D prediction.

PartSTAD. PartSTAD retains the GLIP detector and superpoint representation of PartSLIP, but changes both the 2D region and its contribution to 3D lifting. Each GLIP box is first supplied to SAM (Kirillov et al., 2023) as a spatial prompt, producing a foreground mask with a tighter object-part boundary. Point-to-region membership is therefore evaluated against the SAM mask rather than the full rectangular box. PartSTAD further assigns a positive weight $W(b)$ to every detected region. The weight predictor takes the GLIP box feature together with positional encodings of the camera direction and the 2D box location. All boxes from the same object are processed jointly by a shared two-layer MLP with context normalization, allowing the weight of one region to depend on the complete multi-view detection set. A modified ReLU keeps the predicted weights positive.

Let $\mu_b(\mathbf{p}) \in \{0,1\}$ indicate whether the projection of $\mathbf{p}$ lies inside the SAM mask associated with box b. PartSTAD replaces the unweighted coverage in Eq. 33 with

$$\widetilde{s}_{i,c}^{\mathrm{PST}} = \frac{\sum_{v=1}^{V} \sum_{\mathbf{p}\in\mathcal{P}_i} \nu_v(\mathbf{p}) \max_{b\in\mathcal{B}_{v,c}\cup\{b_\emptyset\}} \big(\mu_b(\mathbf{p}) W(b)\big)}{\sum_{v=1}^{V} \sum_{\mathbf{p}\in\mathcal{P}_i} \nu_v(\mathbf{p})}, \tag{34}$$

where $b_\emptyset$ denotes a dummy empty region with $\mu_{b_\emptyset}(\mathbf{p}) = 0$, providing a zero-valued fallback when no class-c mask covers point $\mathbf{p}$. PartSTAD normalizes the resulting weighted supports over the semantic classes and an additional learnable null state, whose unnormalized score is initialized to 10. Its category-specific weight predictor is trained on eight annotated source shapes using the relaxed 3D mIoU objective, while GLIP and SAM remain frozen.

Within TTA, the trained weight predictor, SAM, and the superpoint lifting pipeline are all fixed. The SAM mask is used as r_i^v, the GLIP part scores provide $\mathbf{s}_i^v$, and the points selected by the mask define $\mathcal{Q}_i^v$. The pretrained PartSTAD weight $W(b)$ remains part of the downstream lifting rule and is not updated by adaptation.

Category-General Source Training and Checkpoint Assignment. The original PartSLIP and PartSTAD implementations organize prompt tuning, data preprocessing, and, for PartSTAD, weight-predictor training around a specified object category. We extend these pipelines to a category-general source-to-target protocol suitable for cross-category evaluation. Let $\mathcal{C}_{\mathrm{s}}$ and $\mathcal{C}_{\mathrm{t}}$ denote the disjoint source and target category sets of a transfer task. For each transfer direction, the annotated shapes

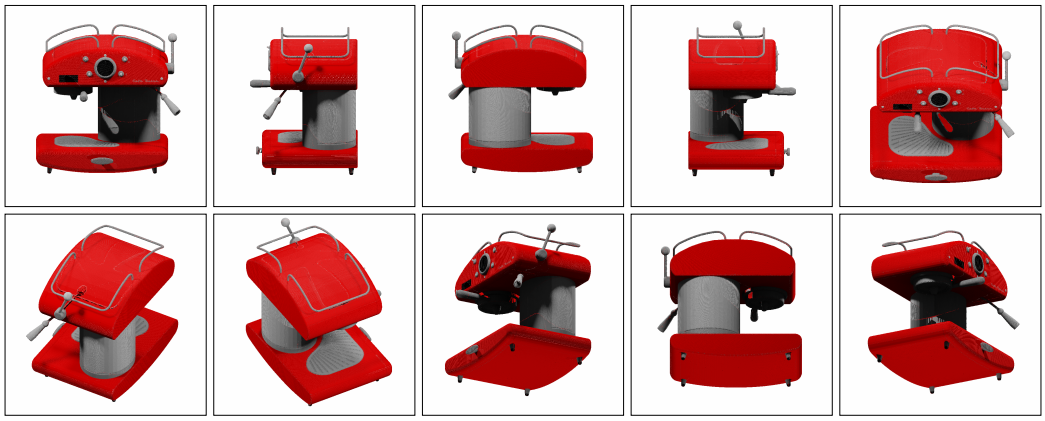

Figure 8: Multi-view renderings of an example point cloud from 10 predefined viewpoints. Each view is rendered at a resolution of 800×800 using the same camera configuration for both base-model branches.

from all categories in $\mathcal{C}_{\mathrm{s}}$ are pooled to train a single source checkpoint, which is subsequently applied to every category in $\mathcal{C}_{\mathrm{t}}$. The training pipeline therefore does not select a separate model according to the target category.

Both base-model branches use the same generic pretrained GLIP checkpoint, and we do not load any GLIP checkpoint fine-tuned for a target category. For PartSTAD, we replace the original category-specific weight predictors with one shared predictor trained jointly on the all source categories. At test time, the target category name and its candidate part names are used solely to construct the open-vocabulary text prompt. The corresponding prompt offset is initialized to zero and adapted using unlabeled target samples, while GLIP, SAM, the source-trained PartSTAD weight predictor, and the lifting modules remain fixed. Consequently, an "unseen target category" denotes a category whose shapes and part annotations are excluded from source training and checkpoint assignment, although its part vocabulary is available as the text query required by the open-vocabulary segmentation task.

D.2 Rendering and Camera Configuration

We use the same multi-view geometry for the two base-model branches so that differences between them arise from their region extraction and lifting rules rather than from camera coverage. Each point cloud is centered and isotropically scaled to fit inside a unit sphere. We render $V = 10$ RGB images with the PyTorch3D point-cloud renderer at a resolution of 800×800, as shown in Fig. 8. The perspective cameras look at the origin from a distance of 2.2. We use a white background, a point radius of 0.005, and one retained point per pixel. The rasterizer additionally returns the visible point index and screen coordinate of every projected point; these correspondences are reused for both region lifting and the construction of $\mathcal{Q}_i^v$.

Following the fixed-view configuration used by PartSLIP and PartSTAD, the ten camera poses are

$$\begin{aligned}\mathcal{V} = \{&(10^\circ, 0^\circ), (10^\circ, 90^\circ), (10^\circ, 180^\circ), (10^\circ, 270^\circ),\\ &(40^\circ, 0^\circ), (40^\circ, 120^\circ), (40^\circ, 240^\circ),\\ &(-20^\circ, 60^\circ), (-20^\circ, 180^\circ), (-20^\circ, 300^\circ)\},\end{aligned} \tag{35}$$

where each ordered pair specifies $(\text{elevation}, \text{azimuth})$. The four near-horizontal views provide full azimuthal coverage, while the remaining upper and lower views expose surfaces that are easily occluded from a single elevation. We keep the ordering and calibration of these cameras fixed throughout adaptation. Consequently, a change in cross-view evidence is caused by the adapted prompt rather than by stochastic camera sampling.

D.3 Prompt Parameterization

Both base models represent the requested semantic vocabulary using a single category-specific text sequence. For an object category o with part names $\{t_c\}_{c=1}^{C}$, the hard prompt concatenates the part names in class-index order and appends the object category:

$$T_o^{\mathrm{hard}} = \texttt{``}t_1\texttt{, }t_2\texttt{, ..., }t_C\texttt{ of a }o\texttt{''}. \tag{36}$$

For example, the prompt for the chair category is ``arm, back, seat, leg, wheel of a chair''. The ordering is kept fixed because GLIP grounds individual spans of the input sequence, and changing the order would alter the correspondence between text spans and semantic classes.

We parameterize the test-time prompt as an additive offset in GLIP's language-feature space, following the prompt-tuning implementation used with PartSLIP and PartSTAD. Let the frozen tokenizer and language encoder map Eq. equation 36 to

$$\mathbf{F}_o = \mathrm{Enc}_{\mathrm{text}}\left(T_o^{\mathrm{hard}}\right) \in \mathbb{R}^{L_o \times d_t}, \tag{37}$$

where L_o is the number of valid text tokens and d_t is the language-feature dimension. We associate each prompt with a learnable offset $\mathbf{\Delta}_o \in \mathbb{R}^{L_o \times d_t}$ and feed

$$\widetilde{\mathbf{F}}_o = \mathbf{F}_o + \mathbf{\Delta}_o \tag{38}$$

to the frozen vision-language fusion layers. This parameterization preserves the original tokens, phrase boundaries, and span-to-class mapping while adapting their continuous representations. It also introduces no additional prompt-length hyperparameter, because the number of learnable vectors is determined by the tokenized hard prompt.

Baseline Adaptation Protocol. The compared TTA methods were originally formulated with different trainable parameter sets and were not uniformly designed for prompt adaptation. For each baseline, we therefore evaluated two implementations: one updating the parameters prescribed by its original formulation and another applying its adaptation objective to the shared GLIP prompt interface in Eq. 38. The prompt-based implementation yielded better overall performance with both PartSLIP and PartSTAD and is consequently used for all baseline results reported in the main paper.

Under this protocol, all methods use the same hard prompts, prompt-offset parameterization, initialization, and update budget. Each baseline nevertheless retains its original adaptation objective and algorithmic mechanism; these components are not replaced by the TRACE objective. The reported comparisons therefore evaluate different adaptation mechanisms within a common trainable parameter space, avoiding confounding effects from unequal parameter capacity or optimization budgets. Likewise, all adaptive variants in Tab. 6 use the same prompt initialization and update budget, while only the indicated compatibility graph, geometric calibration, or uncertainty gate is enabled or removed. The offset is initialized from the prompt state of the source checkpoint. For a checkpoint without source-side prompt tuning, this reduces to a zero offset, such that Eq. 38 initially reproduces the hard-prompt prediction.

TTA methods optimize only $\mathbf{\Delta}_o$ using their own objectives. We use AdamW with a learning rate of 0.07 and a weight decay of 0.25. The GLIP image encoder, language encoder, cross-modal fusion layers, and detection head remain frozen. In the PartSTAD branch, SAM and the pretrained weight-prediction network are also frozen. Thus, both branches expose the same test-time parameter interface, and changes in the final 3D predictions arise solely from the adapted language features processed by the fixed multi-view lifting pipeline.

D.4 Pseudocode of TRACE

Algorithm 1 summarizes the online inference and adaptation procedure of TRACE, following the common online TTA protocol adopted in prior work (Niu et al., 2022; 2023; Liang et al., 2025). For each incoming point cloud, TRACE first obtains view-specific part predictions using the current text prompts and outputs the predictions with the base model. It then constructs a semantic-geometric compatibility graph and calibrates the evidence weights according to geometric consensus. The calibrated evidence is consolidated through the closed-form composition rule to parameterize an uncertainty-aware Dirichlet distribution. For numerical stability, we impose a lower bound of 10^{-4} on the consolidated uncertainty support when constructing the Dirichlet parameters: $\widehat{Z}_{n,\perp} \leftarrow \max\left(\widehat{Z}_{n,\perp}, 10^{-4}\right)$. This clipping prevents a near-zero uncertainty support from producing extreme Dirichlet parameters and anomalously large KL-divergence values or gradients. It is used only for the Dirichlet parameterization and KL regularization; the uncertainty gate is evaluated using the original $\widehat{Z}_{n,\perp}$. The gated objective updates only the learnable text prompts. TRACE then carries them forward to subsequent samples.

Algorithm 1: **T**est-time **R**eliable **A**daptation via **C**hain-of-**E**vidence (TRACE).

Input: Unlabeled point cloud stream $\mathcal{X}$, pretrained 2D foundation model f, learnable text prompts $\mathcal{T}$, viewpoints $\{v\}_{v=1}^{V}$, uncertainty budget m, gate threshold τ, weighting coefficient λ.

Output: Point-wise part labels $\{\hat{\mathbf{y}}_i\}_{X_i \in \mathcal{X}}$.

for $X_i \in \mathcal{X}$ **do**

- Render images $\{I^v\}_{v=1}^{V}$ with predefined camera projections;
- Obtain region-level evidence $\{e_j^v\}_{j=1}^{M_v} \leftarrow f(I^v, \mathcal{T})$; `// Eq. (2)`
- Obtain predictions $\mathbf{z}_i$ using the original 2D-to-3D lifting pipeline;
- Assign point-wise part labels $\hat{y}_{i,n} \leftarrow \arg\max_c z_{i,n}^{\mathrm{adapt},c}$ and output $\hat{\mathbf{y}}_i$; `// Eq. (1)`
- **for** *point* $p_n \in X_i$ *with valid evidence* **do**
  - `// 1. Turning multi-view evidence into reliable chains`
  - Gather valid views $\mathcal{O}_n$;
  - Select the smallest valid region per view covering p_n's projection;
  - Obtain semantic logits $\mathbf{z}_n^v$ and geometric support $\mathcal{Q}_n^v$;
  - Construct the semantic-compatibility graph $\mathcal{G}_n$;
  - Convert logits to non-negative evidence: $\mathbf{e}_n^v \leftarrow \mathrm{softplus}(\mathbf{z}_n^v)$;
  - Assign vertex weights with a uncertainty reserve ; `// Eq. (5)`
  - Compute geometric affinities $\mathbf{G}_n$ from $\{\mathcal{Q}_n^v\}_{v \in \mathcal{O}_n}$; `// Eq. (6)`
  - Obtain the dominant eigenvector $\boldsymbol{\rho}_n$ of $\mathbf{G}_n$;
  - Normalize reliability $\rho_{n,v} \leftarrow \rho_{n,v} / \max_{v' \in \mathcal{O}_n} \rho_{n,v'}$; `// Eq. (7)`
  - Calibrate weights to $\bar{w}_n(v, s)$ by the geometric reliability ; `// Eq. (8)`
  - `// 2. Chain-of-evidence reinforcement`
  - Compute chain supports $\hat{Z}_{n,s}$ via closed-form composition:

$$\frac{\prod_{v \in \mathcal{O}} \bar{w}_v(\perp)}{\sum_{c=1}^{C} \prod_{v \in \mathcal{O}} \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - (C-1) \prod_{v \in \mathcal{O}} \bar{w}_v(\perp)}, s = \perp,$$

$$\frac{\prod_{v \in \mathcal{O}} \big(\bar{w}_v(s) + \bar{w}_v(\perp)\big) - \prod_{v \in \mathcal{O}} \bar{w}_v(\perp)}{\sum_{c=1}^{C} \prod_{v \in \mathcal{O}} \big(\bar{w}_v(c) + \bar{w}_v(\perp)\big) - (C-1) \prod_{v \in \mathcal{O}} \bar{w}_v(\perp)}, s \neq \perp;$$

  - Apply a numerical floor to the uncertainty support: $\widehat{Z}_{n,\perp} \leftarrow \max\Big(\widehat{Z}_{n,\perp}, 10^{-4}\Big)$;
  - Set Dirichlet parameters $\boldsymbol{\alpha}_n$ from evidence chains ; `// Eq. (12)`
  - Select the strongest semantic hypothesis $c_n^\star \leftarrow \arg\max_c \hat{Z}_{n,c}$;
  - Compute the uncertainty-gated adaptation objective

$$\mathcal{L}_{\mathrm{TTA},n} \leftarrow -\mathbb{I}[\hat{Z}_{n,\perp} \leq \tau](\log \mathbb{E}_{\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\alpha}_n)}[p_{c_n^\star}] + \lambda\, D_{\mathrm{KL}}(\mathrm{Dir}(\boldsymbol{\alpha}_n) \,\|\, \mathrm{Dir}(\mathbf{1}))\,);$$

- **end**
- Average the losses over valid points: $\mathcal{L}_{\mathrm{TTA}} \leftarrow \mathrm{mean}_{n:\, \mathcal{O}_n \neq \emptyset} \mathcal{L}_{\mathrm{TTA},n}$;
- Update only the learnable text prompts:

$$\mathcal{T} \leftarrow \mathrm{AdamW}(\mathcal{T}, \nabla_{\mathcal{T}} \mathcal{L}_{\mathrm{TTA}})\,;$$

end