UNITY IN DIVERSITY: LOW-RANK MULTI-RELATIONAL TEST-TIME ADAPTATION FOR 3D POINT CLOUD RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models have achieved strong zero-shot performance in 3D recognition, but their robustness degrades under point cloud corruptions and test-time distribution shifts. Existing point cloud test-time adaptation methods rely heavily on a single relation view, which can be unreliable under diverse corruptions. Thus, the paper highlights the global relational structure among unlabeled target samples. Specifically, we find that corrupted target relations retain compact low-rank structure, which indicates that low-rank modeling can be treated as a structural prior and a complementary view to the raw relation. Besides, when geometry is distorted, feature-space relations can be further captured by semantic relations. Raw, Low-rank, and Semantic relations together could formulate a comprehensive global multi-relational structure. Based on these observations, we propose LoReTTA, a training-free multi-relational target-set test-time adaptation framework that exploits target-set structure without updating the pretrained backbone. LoReTTA constructs multiple relation views from frozen target features and zero-shot predictions, models target dependencies through self-representation, and jointly regularizes the resulting views with a lightweight relation-mode low-rank proximal update. The refined relations are then fused into a sparse consensus graph for residual logit propagation. Experiments across multiple 3D vision-language backbones and corrupted and clean benchmarks demonstrate improved zero-shot robustness and generalization without updating the pretrained backbone. The source code will be publicly released upon acceptance.

1 INTRODUCTION

The rapid development of vision-language models (VLMs) has established a paradigm for open-vocabulary 3D point cloud recognition. Building on large-scale image-text contrastive pretraining (Radford et al., 2021), recent 3D VLMs transfer semantic knowledge to point clouds through multi-view projection, prompt-based adaptation, tri-modal alignment, and scalable 3D representation pretraining (Zhang et al., 2022b; Zhu et al., 2023b; Xue et al., 2023; Liu et al., 2023; Xue et al., 2024; Zhou et al., 2024; Zhu et al., 2023a). Despite their strong zero-shot capability, their performance remains fragile under test-time distribution shifts. Geometric noise, occlusion, non-uniform sampling, density variation, and sensor artifacts can distort point representations and weaken point-text alignment (Sun et al., 2022; Ren et al., 2022), making both target representations and predictions unreliable during inference.

Test-time adaptation (TTA) mitigates distribution shifts by exploiting unlabeled target data during inference (Sun et al., 2020; Wang et al., 2021; Zhang et al., 2022a; Wang et al., 2022; Niu et al., 2022; 2023). Existing TTA methods typically adapt models through entropy minimization, pseudo-label refinement, prompt tuning, or prototype updates, while recent point cloud approaches further exploit masked reconstruction, source-target fusion, and geometric priors (Mirza et al., 2023; Wang et al., 2024; Bahri et al., 2025). For 3D VLMs, PointCache (Sun et al., 2025) and UniAdapter (Tamjidi et al., 2026) further explore training-free adaptation through feature caching and prototype refinement. Despite these advances, existing point cloud TTA methods still rely heavily on a single relation view to characterize target samples, which can be unreliable under diverse corruptions. In particular, feature-space relations are sensitive to geometric distortion, whereas prediction-space re-

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.

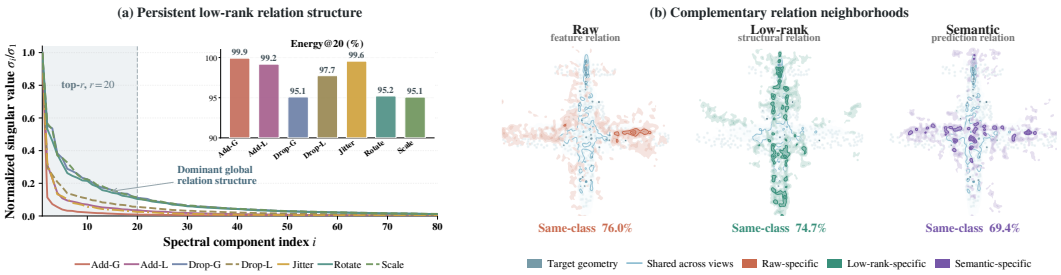

Figure 1: Motivation of LoReTTA. **(a) Corrupted target relations preserve compact low-rank structure.** The leading 20 components retain over 95% of relational energy across seven ModelNet-C corruptions, motivating low-rank modeling as a structural prior. **(b) Multiple relation views provide complementary target dependencies.** Each panel overlays the point clouds of its Top-10 retrieved neighbors. Raw, Low-rank, and Semantic relations induce distinct yet class-consistent neighborhoods, revealing complementary rather than redundant target structure. Dataset-level statistics are reported in Table 1. Ground-truth labels are used only for post-hoc analysis and visualization.

lations may inherit errors from unreliable zero-shot predictions. This motivates more comprehensive modeling of the global relational structure among unlabeled target samples.

We therefore examine the global relational structure of corrupted target sets. First, corrupted target relations retain a compact low-rank structure. As shown in Figure 1(a), the relation spectra decay rapidly across seven ModelNet-C corruptions, with the leading 20 components preserving over 95% of the relational energy. This suggests that low-rank modeling can serve as a structural prior complementary to the fine-grained Raw relation (Liu et al., 2013; Elhamifar & Vidal, 2013; Liu & Yan, 2011; Wang & Xu, 2013). Second, Raw, Low-rank, and Semantic relations capture complementary target dependencies. Figure 1(b) further reveals that, for the same corrupted sample, the three relations recover markedly different yet class-consistent neighborhoods, highlighting complementary rather than redundant dependencies. Table 1 confirms this across ULIP, ULIP-2, and Uni3D. All three relations maintain high same-class purity, while their pairwise Top-10 overlap is only partial. Moreover, Low-rank and Semantic relations recover 28.65%–32.77% of Raw-relation failures, providing complementary evidence beyond the Raw relation. Together, they provide a more comprehensive characterization of the global multi-relational target-set structure.

Based on these observations, we propose LoReTTA, a training-free multi-relational target-set TTA framework that exploits global target-set structure while keeping the pretrained 3D vision-language backbone frozen. To exploit diverse target relations rather than relying on a single potentially unreliable view, LoReTTA constructs complementary Raw, Low-rank, and Semantic relation views from frozen target features and zero-shot predictions, where the Low-rank relation introduces the structural prior and the Semantic relation provides complementary prediction-space evidence. Rather than directly combining pairwise similarities, LoReTTA applies self-representation to each view to capture global dependencies among target samples. The relation-specific coefficient matrices are then stacked into a relation tensor and refined by a lightweight relation-mode low-rank proximal update, which extracts shared cross-view structure while suppressing view-specific variations. Finally, the refined relations are fused into a sparse consensus graph for residual logit propagation, enabling robust prediction refinement without updating the pretrained backbone. Our contributions are summarized as follows:

- We observe that corrupted target relations retain compact low-rank structure, while Raw, Low-rank, and Semantic relations provide complementary target dependencies.
- We propose LoReTTA, a training-free multi-relational target-set TTA framework that combines self-representation, relation-mode low-rank regularization, and consensus-guided prediction refinement.
- LoReTTA performs consistently across diverse corrupted and clean benchmarks, demonstrating strong robustness and generalization. These gains require no updates to the pretrained backbone.

Table 1: **Complementarity of relation views on ModelNet-C.** Results are averaged over seven corruptions at severity level 3. Rescue denotes Raw-relation failures recovered by either the Low-rank or Semantic relation.

Backbone	Same-class purity (%)			Top-10 overlap (%)			Complementary rescue (%)
	Raw relation	Low-rank relation	Semantic relation	Raw–Low-rank	Raw–Semantic	Low-rank–Semantic	
ULIP	75.07	70.99	71.65	70.95	46.95	42.60	28.65
ULIP-2	76.81	72.06	72.99	65.97	45.40	40.53	31.45
Uni3D	81.79	78.48	79.67	61.11	44.22	41.18	32.77

2 Related Work

2.1 Vision-Language Learning for 3D Recognition

Recent 3D vision-language learning has progressively extended language-guided recognition from image-based models to point cloud understanding. Early approaches adapt pretrained vision-language models to 3D data through multi-view projections and prompt-based designs (Zhang et al., 2022b; Zhu et al., 2023b), while subsequent methods learn stronger cross-modal representations through point-image-text alignment and large-scale 3D pretraining (Xue et al., 2023; Liu et al., 2023; Xue et al., 2024; Zhou et al., 2024; Zhu et al., 2023a). These methods primarily improve the representation and alignment capability of 3D VLMs. Our work instead focuses on their test-time robustness, exploiting global dependencies within the unlabeled target set while keeping the pretrained backbone frozen.

2.2 Test-Time Adaptation for Point Clouds

TTA improves model robustness to distribution shifts by exploiting unlabeled target data during inference. Prior work has explored self-supervised objectives, entropy minimization, augmentation consistency, continual adaptation, and reliable sample selection (Sun et al., 2020; Wang et al., 2021; Zhang et al., 2022a; Wang et al., 2022; Niu et al., 2022; 2023). For vision-language models, recent methods further investigate prompt-based adaptation and lightweight training-free refinement (Shu et al., 2022; Karmanov et al., 2024).

For 3D recognition, prior TTA methods exploit masked reconstruction, source-target fusion, and geometric priors (Mirza et al., 2023; Wang et al., 2024; Bahri et al., 2025). PointCache (Sun et al., 2025) and UniAdapter (Tamjidi et al., 2026) further extend training-free adaptation to 3D VLMs through feature caching and prototype refinement. These methods mainly adapt predictions using a dominant feature, prototype, or geometric signal. In contrast, LoReTTA models the unlabeled target set from multiple complementary relational views for prediction refinement.

2.3 Low-Rank Relational Learning

Low-rank modeling provides a classical tool for recovering compact structure from noisy high-dimensional data. Low-rank representation and subspace clustering use self-representation to capture sample dependencies and latent subspace structure (Liu et al., 2013; Elhamifar & Vidal, 2013; Liu & Yan, 2011; Wang & Xu, 2013), while multi-view and tensor-based extensions exploit cross-view correlations to recover shared structure (Zhang et al., 2014).Building on the compact low-rank structure of corrupted target relations, LoReTTA combines self-representation with relation-mode low-rank coupling to reconcile complementary Raw, Low-rank, and Semantic relations for prediction refinement.

3 Method

3.1 Overview

We consider training-free TTA for point cloud recognition with pretrained 3D VLMs. Given an unlabeled target set $\mathcal{X} = \{x_i\}_{i=1}^{N}$ and class prompts $\mathcal{T} = \{t_c\}_{c=1}^{C}$, we keep the pretrained backbone frozen. For each target point cloud x_i, the frozen 3D VLM extracts a point representation $z_i^{\text{raw}} \in \mathbb{R}^D$ and zero-shot logits $y_i \in \mathbb{R}^C$. Despite corruption, target samples retain structured

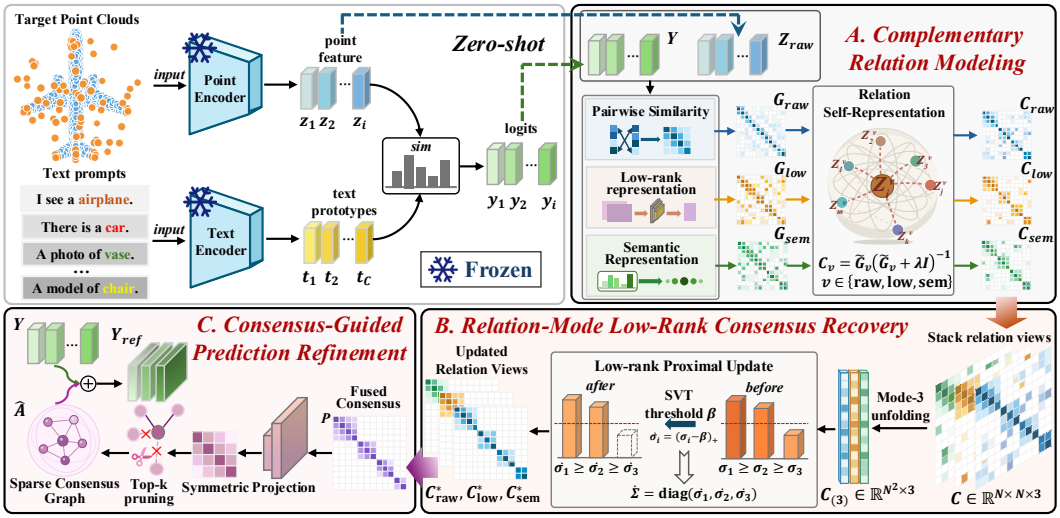

Figure 2: **Overview of LoReTTA.** Given target point clouds and textual class prompts, a frozen 3D VLM extracts target features and zero-shot logits. LoReTTA constructs complementary Raw, Low-rank, and Semantic relations and models their target dependencies through self-representation. A relation-mode low-rank proximal update then captures shared structure across the complementary relation views. Finally, the refined relations are fused into a sparse consensus graph for residual logit propagation and prediction refinement.

set-level dependencies, while no single relation view remains consistently reliable. LoReTTA therefore jointly exploits complementary Raw, Low-rank, and Semantic relations to model the global relational structure of the target set.

As illustrated in Figure 2, LoReTTA consists of three stages. First, *complementary relation modeling* constructs Raw, Low-rank, and Semantic relations from frozen target features and zero-shot predictions, and models their target-set dependencies through self-representation. Second, *relation-mode low-rank consensus recovery* captures shared cross-view structure across the complementary relation views via a low-rank proximal update. Finally, *consensus-guided prediction refinement* fuses the recovered relations into a sparse consensus graph and refines the zero-shot logits through residual propagation. LoReTTA uses frozen representations and predictions without additional training.

3.2 Complementary Relation Modeling

Let $Z_{\text{raw}} = [z_1^{\text{raw}}, \ldots, z_N^{\text{raw}}]^\top \in \mathbb{R}^{N \times D}$ denote the row-normalized target features extracted by the frozen 3D VLM, and $Y = [y_1, \ldots, y_N]^\top \in \mathbb{R}^{N \times C}$ denotes the corresponding zero-shot logits. Under corruption, neither feature nor prediction similarity remains consistently reliable. We therefore describe the target set from three complementary perspectives: Raw, Low-rank, and Semantic. Then they are transformed into target-set dependencies through relation-specific self-representation.

Complementary relation construction. We first construct the *Raw relation* directly from the frozen point-cloud features as $G_{\text{raw}} = Z_{\text{raw}} Z_{\text{raw}}^\top$. It preserves the fine-grained similarity encoded by the pretrained representation and retains rich information from the original feature space. However, its local relations remain sensitive to corruption.

To capture the compact relational structure retained under corruption, we further construct a *Low-rank relation* to emphasize the dominant structure of the target set. Specifically, we decompose $Z_{\text{raw}} = U \Sigma V^\top$ and retain the leading r singular components. Using the spectral coordinates $U_r \Sigma_r$, we define:

$$Z_{\text{low}} = \text{Norm}(U_r \Sigma_r), \qquad G_{\text{low}} = Z_{\text{low}} Z_{\text{low}}^\top. \tag{1}$$

By focusing on the dominant subspace, the Low-rank relation emphasizes dominant global structure while reducing components outside the leading subspace. However, this compression may discard

fine-grained discriminative information preserved by the Raw relation. The two views therefore provide complementary descriptions of the target structure.

Feature-space relations mainly reflect the geometry of target representations, but do not explicitly capture the category-level evidence contained in zero-shot predictions. We therefore construct a *Semantic relation* from the predicted class distributions. Specifically, we compute the class probabilities $P = \mathrm{softmax}(Y/\tau_s)$ and normalize each row as $\bar{P} = \mathrm{Norm}(P)$. The Semantic relation is then defined as

$$G_{\mathrm{sem}} = \bar{P}\bar{P}^{\top}. \tag{2}$$

Unlike the Raw and Low-rank relations, G_{sem} captures prediction-level similarity and can complement feature-space relations when their geometry is distorted, although it may inherit errors from unreliable zero-shot predictions. Together, the three relations provide complementary but imperfect views of the target structure.

Relation-specific self-representation. The above relation matrices characterize pairwise similarities between target samples. However, pairwise affinity alone does not explicitly model how each sample depends on the remaining target set. We further perform self-representation within each relation view, transforming heterogeneous pairwise relations into a common sample-to-sample dependency representation, providing a unified basis for subsequent cross-view coupling.

Before self-representation, we rescale G_{raw} and G_{sem} to match the Frobenius norm of G_{low}, i.e., $\widetilde{G}_v = G_v\|G_{\mathrm{low}}\|_F/\|G_v\|_F$ for $v \in \{\mathrm{raw}, \mathrm{sem}\}$, while $\widetilde{G}_{\mathrm{low}} = G_{\mathrm{low}}$. This places the three relation views on comparable scales under the same regularization strength.

For each relation view $v \in \{\mathrm{raw}, \mathrm{low}, \mathrm{sem}\}$, we compute a ridge-regularized self-representation matrix $C_v \in \mathbb{R}^{N\times N}$:

$$C_v = \arg\min_{C} \frac{1}{2}\,\mathrm{Tr}\left(C\widetilde{G}_v C^{\top}\right) - \mathrm{Tr}\left(C\widetilde{G}_v\right) + \frac{\lambda}{2}\|C\|_F^2 = \widetilde{G}_v(\widetilde{G}_v + \lambda I)^{-1}, \tag{3}$$

where λ controls the ridge regularization strength and $C \in \mathbb{R}^{N\times N}$ denotes the self-representation coefficient matrix. Each row of C_v characterizes the dependency of a target sample on the remaining target set under relation view v. This yields three complementary dependency views, C_{raw}, C_{low}, and C_{sem}, which are subsequently coupled across relation views.

3.3 Relation-Mode Low-Rank Consensus Recovery

The three self-representation matrices provide complementary but imperfect target-dependency views. Direct fusion may preserve view-specific perturbations, while enforcing identical structures may lose complementary information. Therefore, we impose a relation-mode low-rank constraint to promote compact cross-view structure and suppress inconsistencies.

Before cross-view coupling, we normalize each self-representation matrix to remove scale differences. Let $s_v = \|C_v\|_F$ and $\bar{C}_v = C_v/s_v$ for $v \in \{\mathrm{raw}, \mathrm{low}, \mathrm{sem}\}$. The normalized views are stacked and unfolded along the relation mode:

$$C_{(3)} = \left[\mathrm{vec}(\bar{C}_{\mathrm{raw}}), \mathrm{vec}(\bar{C}_{\mathrm{low}}), \mathrm{vec}(\bar{C}_{\mathrm{sem}})\right] \in \mathbb{R}^{N^2\times 3}. \tag{4}$$

Each column preserves the sample-pair dependencies of one relation view, while the three columns jointly expose their cross-view correlations.

To regularize these correlations, we impose a nuclear-norm proximal update on the unfolded representation:

$$C^{\star}_{(3)} = \arg\min_{Q} \frac{1}{2}\|Q - C_{(3)}\|_F^2 + \beta\|Q\|_*, \tag{5}$$

where β controls the relation-mode low-rank regularization. The proximal problem admits a closed-form solution via singular value thresholding. Given $C_{(3)} = U\Sigma V^{\top}$, we obtain $C^{\star}_{(3)} = U\,\mathrm{diag}((\sigma_j - \beta)_+)V^{\top}$, where $(a)_+ = \max(a, 0)$. The resulting columns are reshaped into the three refined relation views and rescaled to their original Frobenius norms. Since $C_{(3)} \in \mathbb{R}^{N^2\times 3}$ has only three columns, the update involves at most three singular values and remains lightweight.

3.4 Consensus-Guided Prediction Refinement

The relation-mode low-rank update yields three refined dependency views with reduced cross-view inconsistencies.We further fuse them into a consensus graph for target-set prediction refinement.

Consensus graph reconstruction. We integrate the refined Raw, Low-rank, and Semantic relation views into a consensus relation through weighted fusion:

$$P = \frac{\alpha_{\mathrm{raw}} C^{\star}_{\mathrm{raw}} + \alpha_{\mathrm{low}} C^{\star}_{\mathrm{low}} + \alpha_{\mathrm{sem}} C^{\star}_{\mathrm{sem}}}{\alpha_{\mathrm{raw}} + \alpha_{\mathrm{low}} + \alpha_{\mathrm{sem}}}, \tag{6}$$

where α_{raw}, α_{low}, and α_{sem} control the contributions of the three views. The resulting P reconciles their complementary target dependencies and provides the basis for constructing the propagation graph. We then project negative affinities to zero, remove self-connections, symmetrize P, retain the top-k neighbors, and row-normalize the result to obtain the sparse consensus graph $\hat{A}$.

Residual prediction refinement. Using the reconstructed consensus graph, we propagate prediction information among related target samples while preserving the original zero-shot predictions:

$$Y_{\mathrm{ref}} = (1 - \eta) Y + \eta \hat{A} Y, \tag{7}$$

where η controls the propagation strength. The residual formulation combines the original zero-shot predictions with relational evidence from the target set, avoiding excessive deviation from the pretrained model and improving robustness.

4 Experiments

4.1 Datasets

We evaluate LoReTTA on three corrupted point cloud benchmarks, ModelNet-C (Ren et al., 2022), ShapeNet-C (Mirza et al., 2023), and ScanObjectNN-C (Wang et al., 2023), covering seven common corruptions: global/local outlier addition, global/local dropout, rotation, scaling, and jittering. Unless otherwise specified, we report results at severity level 3. For ScanObjectNN-C, we evaluate on the Obj_Only, Obj_BG, and Obj_Hardest splits. We further consider Objaverse-LVIS (Deitke et al., 2022) and Omni3D (Brazil et al., 2023) to evaluate LoReTTA on larger and more diverse open-vocabulary 3D recognition benchmarks.

4.2 Experimental Setup

Backbones and baselines. We evaluate LoReTTA on three frozen 3D VLMs: ULIP (Xue et al., 2023), ULIP-2 (Xue et al., 2024), and Uni3D (Zhou et al., 2024). We compare against TDA (Karmanov et al., 2024), CloudFixer (Shim et al., 2024), 3DD-TTA (Dastmalchi et al., 2024), PointCache (Sun et al., 2025), and UniAdapter (Tamjidi et al., 2026). All methods use the same backbones, target data, class prompts, and corruption settings.

Evaluation protocol. LoReTTA performs training-free target-set adaptation. TDA, CloudFixer, and 3DD-TTA use batch size 1, while PointCache and UniAdapter follow their original online protocols. We report top-1 accuracy over seven corruptions at severity level 3 and on clean benchmarks.

Implementation details. We use a fixed configuration for all experiments: $r = 20$, $\lambda = 0.1$, $\beta = 0.05$, $\tau_s = 10$, $k = 10$, $(\alpha_{\mathrm{raw}}, \alpha_{\mathrm{low}}, \alpha_{\mathrm{sem}}) = (0.1, 1.0, 0.01)$, and $\eta = 0.7$. The same hyperparameter configuration is used across all datasets, corruption types, and backbones.

4.3 Main Results

Robustness on corrupted point clouds. As shown in Table 2, LoReTTA achieves the best average accuracy across all three backbones, reaching 52.83%, 62.62%, and 76.22% on ULIP, ULIP-2, and Uni3D, outperforming the strongest baseline by 1.06, 1.51, and 4.66 points, respectively. The gains are particularly evident under dropout, rotation, and scaling corruptions, demonstrating consistent robustness across different 3D VLMs.

Table 2: **Robustness on ModelNet-C.** Top-1 accuracy (%) under seven corruptions at severity level 3. Best results are **bolded** and second-best results are underlined.

Backbone	Method	Corruption							Avg.
		Add-G	Add-L	Drop-G	Drop-L	Rot.	Scale	Jitter	
ULIP (Xue et al., 2023)	Zero-shot	30.55	41.08	50.47	46.20	52.30	47.01	41.80	44.20
	TDA† (Karmanov et al., 2024)	**43.48**	47.93	56.16	53.85	57.86	53.65	49.43	51.77
	CloudFixer† (Shim et al., 2024)	36.95	42.10	46.56	42.91	46.52	40.36	44.65	42.86
	3DD-TTA† (Dastmalchi et al., 2024)	32.54	46.96	54.74	47.85	52.23	46.72	50.45	47.36
	PointCache (Sun et al., 2025)	37.16	43.11	56.44	50.93	57.62	53.40	26.01	46.38
	UniAdapter (Tamjidi et al., 2026)	17.22	51.13	54.86	**56.85**	**58.75**	53.24	49.23	48.76
	LoReTTA	38.13	**51.30**	**58.23**	55.31	58.71	**55.51**	**52.59**	**52.83**
ULIP-2 (Xue et al., 2024)	Zero-shot	63.25	51.38	61.87	53.74	68.11	66.65	15.94	54.42
	TDA† (Karmanov et al., 2024)	67.99	59.56	70.79	60.98	72.00	68.44	28.04	61.11
	CloudFixer† (Shim et al., 2024)	41.37	35.01	39.42	33.10	45.22	39.42	26.94	37.21
	3DD-TTA† (Dastmalchi et al., 2024)	53.97	49.35	57.01	46.60	55.71	54.94	**44.17**	51.68
	PointCache (Sun et al., 2025)	65.68	54.94	69.00	58.67	70.87	67.14	22.93	58.46
	UniAdapter (Tamjidi et al., 2026)	69.98	**62.48**	38.33	56.81	73.70	71.56	27.03	57.13
	LoReTTA	**70.34**	62.20	**72.41**	**62.88**	**74.07**	**72.45**	23.99	**62.62**
Uni3D (Zhou et al., 2024)	Zero-shot	71.10	56.06	66.31	65.08	78.47	73.24	43.92	64.88
	TDA† (Karmanov et al., 2024)	70.58	67.59	68.96	65.28	79.90	75.53	63.41	70.18
	CloudFixer† (Shim et al., 2024)	62.20	54.46	47.73	46.92	61.87	55.02	55.06	54.75
	3DD-TTA† (Dastmalchi et al., 2024)	58.71	56.32	71.03	59.64	69.08	64.22	**66.05**	63.58
	PointCache (Sun et al., 2025)	73.15	69.47	68.29	67.90	79.42	74.54	60.24	70.43
	UniAdapter (Tamjidi et al., 2026)	**83.75**	**76.99**	52.23	63.78	86.43	81.16	56.60	71.56
	LoReTTA	78.16	73.82	**80.27**	**79.09**	**88.01**	**83.14**	51.05	**76.22**

† Adapted baselines evaluated with the same frozen 3D-VLM backbones and evaluation data.

Table 3: **Robustness on ScanObjectNN-C_Obj_Hardest.** Top-1 accuracy (%) under seven corruptions at severity level 3.

Backbone	Method	Corruption							Avg.
		Add-G	Add-L	Drop-G	Drop-L	Rot.	Scale	Jitter	
ULIP (Xue et al., 2023)	Zero-shot	30.60	30.67	30.92	23.77	26.82	26.27	21.79	27.26
	TDA† (Karmanov et al., 2024)	32.82	32.96	32.41	27.52	28.24	27.31	23.18	29.21
	CloudFixer† (Shim et al., 2024)	25.78	25.78	26.44	20.78	22.31	21.17	21.96	23.46
	3DD-TTA† (Dastmalchi et al., 2024)	28.35	28.04	26.75	19.60	24.08	24.22	22.59	24.80
	PointCache (Sun et al., 2025)	32.41	31.78	32.10	26.27	28.70	27.55	22.90	28.82
	UniAdapter (Tamjidi et al., 2026)	21.65	18.04	**35.84**	**29.70**	**30.46**	27.38	**23.91**	26.71
	LoReTTA	**34.66**	**34.77**	34.70	24.95	29.74	**28.70**	22.38	**29.99**
ULIP-2 (Xue et al., 2024)	Zero-shot	34.14	33.45	28.45	24.88	28.45	30.05	12.77	27.46
	TDA† (Karmanov et al., 2024)	**39.21**	**38.34**	**33.07**	**30.40**	**31.78**	**32.79**	18.01	**31.94**
	CloudFixer† (Shim et al., 2024)	22.59	22.21	22.21	18.36	20.82	20.71	18.39	20.76
	3DD-TTA† (Dastmalchi et al., 2024)	26.41	26.47	21.51	19.08	20.75	21.89	**22.03**	22.59
	PointCache (Sun et al., 2025)	36.81	36.54	31.05	27.69	30.36	32.06	14.64	29.88
	UniAdapter (Tamjidi et al., 2026)	31.78	28.35	29.94	25.05	30.08	30.85	13.91	27.14
	LoReTTA	35.95	35.18	28.59	23.80	29.56	31.12	12.60	28.11
Uni3D (Zhou et al., 2024)	Zero-shot	42.30	42.02	32.58	29.39	39.97	35.39	25.99	35.38
	TDA† (Karmanov et al., 2024)	43.51	44.00	35.22	32.41	41.22	36.99	**28.42**	37.40
	CloudFixer† (Shim et al., 2024)	30.99	31.37	26.13	19.88	29.39	25.29	27.90	27.28
	3DD-TTA† (Dastmalchi et al., 2024)	33.14	33.17	29.60	21.44	28.70	26.20	25.05	28.19
	PointCache (Sun et al., 2025)	46.53	46.56	38.83	31.92	44.00	39.10	20.47	38.20
	UniAdapter (Tamjidi et al., 2026)	46.36	41.53	**43.48**	**40.35**	45.73	**42.19**	19.36	39.86
	LoReTTA	**50.03**	**50.62**	41.57	32.03	**47.09**	41.22	18.81	**40.20**

† Adapted baselines evaluated with the same frozen 3D-VLM backbones and evaluation data.

Robustness on real-world scanned objects. Table 3 reports results on ScanObjectNN-C_Obj_Hardest. LoReTTA achieves the best average accuracy with ULIP and Uni3D, reaching 29.99% and 40.20%, respectively, outperforming the strongest baselines by 0.78 and 0.34 percentage points. On ULIP-2, LoReTTA reaches 28.11%, below TDA at 31.94%, but still slightly improves the zero-shot baseline of 27.46%. The weaker gains mainly appear under Drop-L and Jitter, where severe geometric corruption can reduce the reliability of target relations and thus limit relational

Table 4: Clean-set generalization.

Dataset	ULIP				ULIP-2				Uni3D			
	Zero-shot	PointCache	UniAdapter	**LoReTTA**	Zero-shot	PointCache	UniAdapter	**LoReTTA**	Zero-shot	PointCache	UniAdapter	**LoReTTA**
ModelNet	56.89	**64.22**	58.23	59.48	71.23	74.53	**76.47**	74.58	78.84	83.87	83.96	**86.91**
ScanObj	47.68	49.44	49.57	**54.22**	42.00	43.73	40.54	**43.89**	56.63	57.16	**60.07**	59.72
OLVIS	6.08	**7.02**	6.49	6.33	30.26	**32.36**	31.83	31.22	46.20	47.13	47.49	**50.27**
Omni3D	8.39	**10.47**	10.05	10.25	26.36	29.37	**29.58**	28.61	32.06	33.22	35.95	**36.13**
Avg.	29.76	**32.79**	31.09	32.57	42.46	**45.00**	44.61	44.58	53.43	55.35	56.87	**58.26**

Backbones: ULIP (Xue et al., 2023), ULIP-2 (Xue et al., 2024), and Uni3D (Zhou et al., 2024). Baselines: PointCache (Sun et al., 2025) and UniAdapter (Tamjidi et al., 2026).

Table 5: Component ablation.

Variant	ULIP	ULIP-2	Uni3D
LR affinity	50.10	59.39	72.05
Raw SR	50.91	61.03	73.32
+Low-rank SR	51.24	61.39	73.91
+Semantic SR	51.54	61.91	75.37
+Relation fusion	51.72	62.05	75.89
LoReTTA	**52.83**	**62.62**	**76.22**

Table 6: Fusion strategy ablation.

Method	ULIP	ULIP-2	Uni3D
Concat	51.51	60.68	73.55
Average	51.29	62.12	74.18
Weighted sum	51.79	62.35	75.89
Independent LR	52.29	61.05	75.02
Relation-mode LR	**52.83** (+0.54)	**62.62** (+1.57)	**76.22** (+1.20)

propagation. Overall, LoReTTA remains effective across challenging scanned-object corruptions, particularly with ULIP and Uni3D.

Generalization on clean target datasets. Table 4 evaluates LoReTTA on clean target sets. LoReTTA remains competitive across three backbones and achieves the best average performance with Uni3D at 58.26%. It also shows strong gains on ScanObjectNN and the large-scale OLVIS and Omni3D benchmarks with Uni3D. Although slightly below the strongest baselines on ULIP and ULIP-2, LoReTTA maintains robust clean-set generalization without corruption-specific tuning.

4.4 Ablation Study

Effect of each component. Table 5 shows consistent improvements as LoReTTA progressively incorporates its main components. Self-representation already improves over pairwise low-rank affinity, while adding the Low-rank and Semantic relations further strengthens performance across all three backbones. Relation fusion reaches 51.72%, 62.05%, and 75.89%, and the relation-mode low-rank consensus further improves these results to 52.83%, 62.62%, and 76.22%, respectively. These results verify the complementary benefits of multi-relational modeling and cross-view consensus recovery.

Relation-mode low-rank consensus. Table 6 compares different strategies for integrating relation views. Independent low-rank processing is not consistently better than simple fusion. Relation-mode low-rank coupling performs best on all three backbones, improving over independent processing by 0.54, 1.57, and 1.20 points on ULIP, ULIP-2, and Uni3D, respectively.This highlights the benefit of jointly modeling shared structure across relation views.

Rank sensitivity. Figure 3 evaluates the retained rank on ModelNet-C with ULIP, ULIP-2, and Uni3D. All three backbones perform best at $r = 20$, reaching 52.83%, 62.62%, and 76.22%, respectively. Lower ranks may discard useful target structure, whereas higher ranks may retain corruption-specific variation, suggesting that a moderate-rank subspace offers a favorable trade-off.

Target-set scalability. Table 7 shows that LoReTTA benefits from larger target sets across all three backbones. Performance improves substantially as more target samples become available, while the adaptation overhead remains modest, requiring at most 0.126 s and 0.62 GB of peak GPU memory at full scale. The dense $N \times N$ relation matrices require $\mathcal{O}(N^2)$ memory, which may limit scalability to much larger target sets, while the relation-mode update itself remains lightweight with only three relation views.

4.5 Visualization

Discriminative structure concentration. Figure 4 shows that the dominant rank-20 subspace preserves clear class-wise structure, whereas the discarded spectral residual is substantially more mixed.

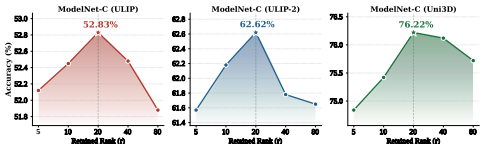

Figure 3: **Rank sensitivity.** Performance consistently peaks at a moderate rank ($r = 20$).

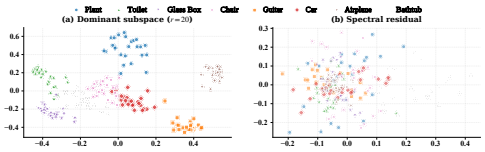

Figure 4: **Discriminative structure concentration.** Dominant components retain class structure, while residuals are mixed.

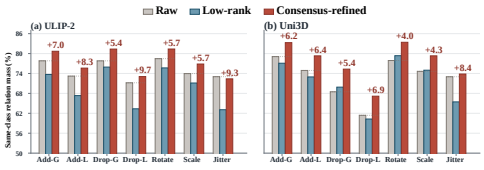

Figure 5: **Relation recovery under corruptions.** Consensus strengthens same-class connectivity.

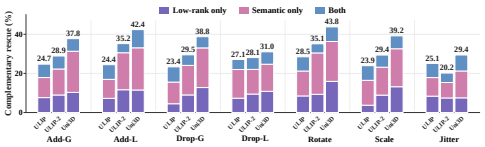

Figure 6: **Complementary rescue across corruptions.** Low-rank and Semantic relations provide complementary rescue.

Table 7: **Target-set scalability of LoReTTA on ModelNet-C.** Accuracy is reported in %, runtime in seconds, and peak memory in GB. Results for $N <$ Full ($N = 2468$) are averaged over three subset seeds.

	ULIP			ULIP-2			Uni3D		
N	Acc.	Time	Mem.	Acc.	Time	Mem.	Acc.	Time	Mem.
256	45.98 $\pm$ 2.11	0.0183 $\pm$ 0.0039	0.0080	57.94 $\pm$ 2.23	0.0311 $\pm$ 0.0282	0.0080	70.70 $\pm$ 1.89	0.0257 $\pm$ 0.0091	0.0094
512	49.93 $\pm$ 2.16	0.0566 $\pm$ 0.0162	0.0286	59.43 $\pm$ 1.50	0.0806 $\pm$ 0.0106	0.0286	73.08 $\pm$ 0.92	0.0903 $\pm$ 0.0847	0.0315
1024	51.60 $\pm$ 0.68	0.0616 $\pm$ 0.0230	0.1050	60.00 $\pm$ 0.38	0.0811 $\pm$ 0.0225	0.1050	75.47 $\pm$ 0.82	0.1135 $\pm$ 0.0253	0.1138
Full	**52.83**	0.0825	0.5945	**62.62**	0.1083	0.5945	**76.22**	0.1259	0.6123

This indicates that discriminative target structure is primarily concentrated in the leading spectral components, while the residual exhibits substantially weaker class separability.

Relation recovery visualization. Figure 5 shows that relation-mode consensus consistently recovers same-class connectivity degraded by low-rank projection across corruptions and backbones, supporting low-rank structure as a shared prior rather than the final relation.

Complementary rescue across corruptions. Figure 6 shows that Low-rank and Semantic relations independently recover Raw-relation failures across corruptions and backbones, while their shared rescues capture additional consistent dependencies. This complementary yet overlapping evidence further supports joint multi-relational modeling in LoReTTA.

5 Conclusion

We presented LoReTTA, a training-free multi-relational target-set TTA framework for robust 3D point cloud recognition with pretrained VLMs. Our study reveals that corrupted target relations retain compact low-rank structure, while Raw, Low-rank, and Semantic relations provide complementary target dependencies. Building on these observations, LoReTTA models target-set dependencies through self-representation and captures shared cross-view structure with a lightweight relation-mode low-rank proximal update. The resulting relations are fused into a sparse consensus graph for residual prediction refinement without updating the pretrained backbone. Experiments across multiple 3D VLMs and corrupted and clean benchmarks demonstrate consistent improvements in robustness and generalization. Overall, LoReTTA highlights the potential of jointly exploiting diverse relational evidence and shared low-rank structure for reliable TTA in 3D recognition.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, grammatical checking, code debugging, and LaTeX formatting. All references were independently searched and verified by the authors, and all research ideas, experimental settings, results, analyses, and conclusions were determined and reviewed by the authors.

REFERENCES

Ali Bahri, Moslem Yazdanpanah, Sahar Dastani, Mehrdad Noori, Gustavo Adolfo Vargas Hakim, David Osowiechi, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. SMART-PC: Skeletal model adaptation for robust test-time training in point clouds. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2442–2457. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bahri25a.html.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13154–13164. IEEE, 2023.

Hamidreza Dastmalchi, Aijun An, Ali Cheraghian, Shafin Rahman, and Sameera Ramasinghe. Test-time adaptation of 3d point clouds via denoising diffusion models, 2024. URL https://arxiv.org/abs/2411.14495.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects, 2022. URL https://arxiv.org/abs/2212.08051.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. doi: 10.1109/TPAMI.2013.57.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171, June 2024.

Guangcan Liu and Shuicheng Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *2011 International Conference on Computer Vision*, pp. 1615–1622, 2011. doi: 10.1109/ICCV.2011.6126422.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35 (1):171–184, January 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.88. URL https://doi.org/10.1109/TPAMI.2012.88.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023.

M. Jehanzeb Mirza, Inkyu Shin, Wei Lin, Andreas Schriebl, Kunyang Sun, Jaesung Choe, Horst Possegger, Mateusz Kozinski, In So Kweon, Kun-Jin Yoon, and Horst Bischof. Mate: Masked autoencoders are online 3d test-time learners, 2023. URL https://arxiv.org/abs/2211.11432.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world, 2023. URL https://arxiv.org/abs/2302.12400.

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

Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International conference on machine learning*, pp. 18559–18575. PMLR, 2022.

Hajin Shim, Changhun Kim, and Eunho Yang. Cloudfixer: Test-time adaptation for 3d point clouds via diffusion-guided geometric transformation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXI*, pp. 454–471, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72750-4. doi: 10.1007/978-3-031-72751-1_26. URL https://doi.org/10.1007/978-3-031-72751-1_26.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Hongyu Sun, Qiuhong Ke, Ming Cheng, Yongcai Wang, Deying Li, Chenhui Gou, and Jianfei Cai. Point-cache: Test-time dynamic and hierarchical cache for robust and generalizable point cloud analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1263–1275. IEEE, 2025.

Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z. Morley Mao. Benchmarking robustness of 3d point cloud recognition against common corruptions, 2022. URL https://arxiv.org/abs/2201.12296.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Mehran Tamjidi, Hamidreza Dastmalchi, Mohammadreza Alimoradijazi, Ali Cheraghian, Aijun An, and Morteza Saberi. Adapt-as-you-walk through the clouds: training-free online test-time adaptation of 3d vision-language foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9296–9304, 2026.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://api.semanticscholar.org/CorpusID:232278031.

Jie Wang, Lihe Ding, Tingfa Xu, Shaocong Dong, Xinli Xu, Long Bai, and Jianan Li. Sample-adaptive augmentation for point cloud recognition against real-world corruptions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14284–14293. IEEE, 2023.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201, 2022. doi: 10.1109/CVPR52688.2022.00706.

Yanshuo Wang, Ali Cheraghian, Zeeshan Hayder, Jie Hong, Sameera Ramasinghe, Shafin Rahman, David Ahmedt-Aristizabal, Xuesong Li, Lars Petersson, and Mehrtash Harandi. Backpropagation-free network for 3d test-time adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23231–23241, 2024. doi: 10.1109/CVPR52733.2024.02192.

Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 89–97, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/wang13.html.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1189. IEEE, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27081–27091. IEEE, 2024.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: test time robustness via adaptation and augmentation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022a. Curran Associates Inc. ISBN 9781713871088.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8542–8552. IEEE, 2022b.

Zemin Zhang, Gregory Ely, Shuchin Aeron, Ning Hao, and Misha Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-svd, 2014. URL https://arxiv.org/abs/1407.1785.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations*, volume 2024, pp. 46766–46782, 2024.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zong-Rui Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *ArXiv*, abs/2310.01852, 2023a. URL https://api.semanticscholar.org/CorpusID:263608698.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2639–2650. IEEE, 2023b.

A Appendix

A.1 Additional Method Details

Algorithmic overview. Algorithm 1 summarizes the complete inference procedure of LoReTTA. Given an unlabeled target set, LoReTTA keeps the pretrained 3D-VLM fully frozen and first extracts target features and zero-shot predictions. It then constructs three complementary relation views from the raw feature, low-rank structural, and semantic prediction spaces, and obtains a self-representation matrix for each view. These relation matrices are jointly regularized through a relation-mode low-rank proximal update to recover their shared consensus structure. Finally, the refined relations are fused into a sparse target graph and used for residual logit propagation. The entire procedure is training-free and requires no gradient-based parameter updates.

A.1.1 Closed-Form Solution of the Relation-Mode Proximal Subproblem

We provide the closed-form solution of the relation-mode proximal subproblem used in the consensus recovery stage of LoReTTA. Given the normalized relation tensor $\mathcal{C}$, we consider

$$\mathcal{Q}^\star = \arg\min_{\mathcal{Q}} \frac{1}{2} \|\mathcal{Q} - \mathcal{C}\|_F^2 + \beta \|\mathrm{Unfold}_3(\mathcal{Q})\|_* , \tag{8}$$

where $\|\cdot\|_*$ denotes the nuclear norm and $\beta > 0$ controls the amount of low-rank shrinkage along the relation mode.

Throughout this paper, we define the relation-mode unfolding by stacking the vectorized relation slices as columns:

$$C_{(3)} = \mathrm{Unfold}_3(\mathcal{C}) = \left[\mathrm{vec}(\bar{C}_{\mathrm{raw}}), \mathrm{vec}(\bar{C}_{\mathrm{low}}), \mathrm{vec}(\bar{C}_{\mathrm{sem}})\right] \in \mathbb{R}^{N^2 \times 3}, \qquad Q_{(3)} = \mathrm{Unfold}_3(\mathcal{Q}). \tag{9}$$

Under tensor conventions in which the conventional mode-3 unfolding has shape $3 \times N^2$, our definition corresponds to its transpose. This convention does not affect the following analysis, since both the Frobenius norm and the nuclear norm are invariant to transposition. Moreover, the relation-mode unfolding is a bijective rearrangement of tensor entries and therefore preserves the Frobenius norm. Hence, Eq. (8) is equivalent to

$$Q_{(3)}^\star = \arg\min_{Q} \frac{1}{2} \left\|Q - C_{(3)}\right\|_F^2 + \beta \|Q\|_*. \tag{10}$$

Proposition 1. Let the singular value decomposition of $C_{(3)}$ be

$$C_{(3)} = U_c \,\mathrm{diag}(\sigma_1, \ldots, \sigma_p) V_c^\top, \tag{11}$$

where $p = \min\{N^2, 3\} = 3$ and $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$. Then Eq. (10) admits a unique global minimizer given by

$$Q_{(3)}^\star = U_c \,\mathrm{diag}\left((\sigma_1 - \beta)_+, \ldots, (\sigma_p - \beta)_+\right) V_c^\top, \tag{12}$$

where $(x)_+ = \max(x, 0)$.

Proof. Define

$$F(Q) = \frac{1}{2}\|Q - C_{(3)}\|_F^2 + \beta \|Q\|_*. \tag{13}$$

The quadratic term is 1-strongly convex with respect to Q, while the nuclear norm is convex. Hence, $F(Q)$ is strongly convex and therefore admits a unique global minimizer.

Let the singular value decomposition of a candidate matrix Q be

$$Q = \widetilde{U} \,\mathrm{diag}(s_1, \ldots, s_p) \widetilde{V}^\top, \qquad s_j \geq 0. \tag{14}$$

Expanding the quadratic term yields

$$F(Q) = \frac{1}{2}\|Q\|_F^2 + \frac{1}{2}\|C_{(3)}\|_F^2 - \langle Q, C_{(3)} \rangle + \beta \|Q\|_*. \tag{15}$$

For fixed singular values $\{s_j\}$, minimizing $F(Q)$ is equivalent to maximizing $\langle Q, C_{(3)} \rangle$. By von Neumann's trace inequality,

$$\langle Q, C_{(3)} \rangle \leq \sum_{j=1}^{p} s_j \sigma_j, \tag{16}$$

and the upper bound is attained when the left and right singular vectors of Q are aligned with those of $C_{(3)}$. Consequently, an optimal solution can be written as

$$Q = U_c \operatorname{diag}(s_1, \ldots, s_p) V_c^\top. \tag{17}$$

Substituting this form into Eq. (15) reduces the matrix optimization to p independent scalar problems:

$$s_j^\star = \arg\min_{s_j \geq 0} \frac{1}{2}(s_j - \sigma_j)^2 + \beta s_j. \tag{18}$$

Each scalar objective is strictly convex, and its constrained minimizer is

$$s_j^\star = \max(\sigma_j - \beta, 0) = (\sigma_j - \beta)_+. \tag{19}$$

Substituting the optimal singular values back into the aligned singular vectors yields Eq. (12). Since the objective is strongly convex, this solution is the unique global minimizer.

Implication for LoReTTA. Proposition 1 shows that the relation-mode proximal subproblem does not require an iterative solver. In particular, a single singular value thresholding (SVT) operation exactly evaluates the corresponding nuclear-norm proximal operator. This differs from iterative schemes such as ADMM, which would require repeated primal and dual updates to solve the same proximal subproblem numerically. In LoReTTA, the unfolded relation matrix has only three columns, corresponding to the Raw, Low-rank, and Semantic relation views, making this SVT step particularly lightweight.

After the proximal update, LoReTTA deterministically rescales each recovered relation slice to its pre-consensus Frobenius norm before weighted fusion. This scale-restoration step is a subsequent normalization operation and is not part of the optimization problem in Eq. (8).

A.2 Additional Experimental Results

Results on ShapeNet-C. As shown in Table 8, LoReTTA achieves the best average accuracy across all three backbones, reaching 89.33%, 85.55%, and 94.43% with ULIP, ULIP-2, and Uni3D, respectively. Compared with the strongest baseline for each backbone, the improvements are 0.61, 1.97, and 1.17 percentage points. The gains are observed across diverse corruption types, indicating that the proposed relational adaptation generalizes consistently across different 3D-VLM backbones.

Results on ScanObjectNN-C_Obj_Only. Table 9 reports results on the more challenging Obj_Only setting. LoReTTA achieves the highest average accuracy with ULIP (47.77%), improving the strongest baseline by 0.93 points, and remains competitive with Uni3D. The performance is less consistent with ULIP-2 and Uni3D, where TDA obtains higher averages, suggesting that severe real-world object variations can weaken the reliability of target-set relational structure.

Results on ScanObjectNN-C_Obj_BG. On Obj_BG (Table 10), LoReTTA achieves the best average performance with ULIP and Uni3D, obtaining 43.10% and 53.84%, respectively. With ULIP-2, LoReTTA remains competitive but trails TDA. Overall, these results show that LoReTTA provides robust improvements across most backbone–corruption combinations, while its effectiveness can decrease when corruption substantially disrupts the target-domain neighborhood structure.

A.2.1 Optimization Efficiency of the Relation-Mode Solver

Proximal SVT vs. ADMM. Figure 7 compares the closed-form proximal SVT update with ADMM under different iteration budgets. Across ULIP, ULIP-2, and Uni3D, all solvers achieve the same downstream accuracy, whereas proximal SVT requires substantially less relation-mode solver time. These results complement the analysis in Appendix A.1.1: since the proximal subproblem admits an exact closed-form solution via singular value thresholding, iterative optimization with ADMM is unnecessary and introduces additional computation without improving recognition performance.

Algorithm 1 LoReTTA

Input: Target set $\mathcal{X}$; frozen 3D-VLM; class prompts $\mathcal{T}$; hyperparameters Θ.
Output: Refined logits Y_{ref}.

Extract target features Z and zero-shot logits Y; $Z_{\mathrm{raw}} \leftarrow \mathrm{Norm}(Z)$.
Stage 1: Complementary Relation Modeling
$G_{\mathrm{raw}} \leftarrow Z_{\mathrm{raw}} Z_{\mathrm{raw}}^{\top}, \quad Z_{\mathrm{low}} \leftarrow \mathrm{Norm}(U_r \Sigma_r), \quad G_{\mathrm{low}} \leftarrow Z_{\mathrm{low}} Z_{\mathrm{low}}^{\top}$.
$P \leftarrow \mathrm{softmax}(Y/\tau_s), \quad \bar{P} \leftarrow \mathrm{Norm}(P), \quad G_{\mathrm{sem}} \leftarrow \bar{P}\bar{P}^{\top}$.
$\widetilde{G}_{\mathrm{low}} \leftarrow G_{\mathrm{low}}; \quad \widetilde{G}_v \leftarrow G_v \frac{\|G_{\mathrm{low}}\|_F}{\|G_v\|_F}, \ \ v \in \{\mathrm{raw}, \mathrm{sem}\}$.
for $v \in \{\mathrm{raw}, \mathrm{low}, \mathrm{sem}\}$ **do**
  $C_v \leftarrow \widetilde{G}_v (\widetilde{G}_v + \lambda I)^{-1}, \quad s_v \leftarrow \|C_v\|_F, \quad \bar{C}_v \leftarrow C_v / s_v$.
end for
Stage 2: Relation-Mode Low-Rank Consensus Recovery
$C_{(3)} \leftarrow [\mathrm{vec}(\bar{C}_{\mathrm{raw}}), \mathrm{vec}(\bar{C}_{\mathrm{low}}), \mathrm{vec}(\bar{C}_{\mathrm{sem}})]$.
$C_{(3)} = U_c \Sigma_c V_c^{\top}; \quad C^{\star}_{(3)} \leftarrow U_c \, \mathrm{diag}((\sigma_j - \beta)_+) \, V_c^{\top}$.
Fold $C^{\star}_{(3)}$ and restore scales: $C^{\star}_v \leftarrow s_v \bar{C}^{\star}_v, \ \ \forall v$.
Stage 3: Consensus-Guided Prediction Refinement
$C_{\mathrm{fused}} \leftarrow \frac{\alpha_{\mathrm{raw}} C^{\star}_{\mathrm{raw}} + \alpha_{\mathrm{low}} C^{\star}_{\mathrm{low}} + \alpha_{\mathrm{sem}} C^{\star}_{\mathrm{sem}}}{\alpha_{\mathrm{raw}} + \alpha_{\mathrm{low}} + \alpha_{\mathrm{sem}}}$.
Construct $\hat{A}$ via nonnegative projection, symmetrization, Top-k sparsification, and row normalization.
$Y_{\mathrm{ref}} \leftarrow (1-\eta) Y + \eta \hat{A} Y$.
return Y_{ref}.

Table 8: **Robustness on ShapeNet-C.** Top-1 accuracy (%) under seven corruptions at severity level 3. Best results are **bolded** and second-best results are underlined.

Backbone	Method	Corruption							Avg.
		Add-G	Add-L	Drop-G	Drop-L	Rot.	Scale	Jitter	
ULIP (Xue et al., 2023)	Zero-shot	60.72	77.28	99.03	90.15	97.46	95.34	78.74	85.53
	TDA† (Karmanov et al., 2024)	**77.07**	79.40	98.99	92.00	97.77	95.44	80.38	88.72
	CloudFixer† (Shim et al., 2024)	56.26	80.17	92.73	78.32	86.26	83.72	82.25	79.96
	3DD-TTA† (Dastmalchi et al., 2024)	62.70	83.05	97.70	87.82	94.57	92.59	**90.22**	86.95
	PointCache (Sun et al., 2025)	67.43	77.91	98.96	90.50	96.56	93.44	80.13	86.42
	UniAdapter (Tamjidi et al., 2026)	28.25	75.61	92.35	93.35	92.10	85.69	72.51	77.13
	LoReTTA	67.05	**84.62**	**99.13**	**95.53**	**98.92**	**98.47**	81.59	**89.33**
ULIP-2 (Xue et al., 2024)	Zero-shot	87.93	82.33	93.47	85.49	92.39	90.72	37.24	81.37
	TDA† (Karmanov et al., 2024)	90.19	85.91	94.64	86.78	93.49	92.03	42.00	83.58
	CloudFixer† (Shim et al., 2024)	61.59	54.84	78.46	56.37	68.02	67.29	51.91	62.64
	3DD-TTA† (Dastmalchi et al., 2024)	80.24	75.99	91.02	79.71	84.76	86.60	**80.65**	82.71
	PointCache (Sun et al., 2025)	88.34	83.92	94.57	85.21	93.35	91.86	40.68	82.56
	UniAdapter (Tamjidi et al., 2026)	88.31	84.31	93.46	**91.20**	92.24	90.29	27.21	81.00
	LoReTTA	**91.54**	**88.48**	**96.87**	89.67	**96.17**	**96.07**	40.08	**85.55**
Uni3D (Zhou et al., 2024)	Zero-shot	90.88	93.26	96.25	94.48	95.29	94.52	68.44	90.45
	TDA† (Karmanov et al., 2024)	91.41	91.34	96.24	94.85	97.11	95.93	85.91	93.26
	CloudFixer† (Shim et al., 2024)	88.41	88.10	90.22	84.03	90.92	87.79	**88.83**	88.33
	3DD-TTA† (Dastmalchi et al., 2024)	82.57	76.20	89.14	84.97	87.86	86.57	84.69	84.57
	PointCache (Sun et al., 2025)	91.75	90.92	94.24	94.75	95.04	94.89	82.88	92.07
	UniAdapter (Tamjidi et al., 2026)	96.21	94.43	96.42	95.23	96.62	95.79	74.01	92.67
	LoReTTA	**97.49**	**97.25**	**98.92**	**97.91**	**98.82**	**98.57**	72.06	**94.43**

† Adapted baselines evaluated with the same frozen 3D-VLM backbones and evaluation data.

Detailed scalability results. Table 11 provides the full scalability results corresponding to the analysis in the main paper. For $N <$ Full, results are averaged over three randomly sampled target subsets and reported as mean±standard deviation. Increasing the target-set size generally improves accuracy and reduces performance variance across ULIP, ULIP-2, and Uni3D. At full scale, LoReTTA reaches 52.83%, 62.62%, and 76.22% accuracy, respectively, with less than 0.13 s solver time and 0.62 GB peak GPU memory.

Table 9: **Robustness on ScanObjectNN-C_Obj_Only.** Top-1 accuracy (%) under seven corruptions at severity level 3. Best results are **bolded** and second-best results are underlined.

Backbone	Method	Corruption							Avg.
		Add-G	Add-L	Drop-G	Drop-L	Rot.	Scale	Jitter	
ULIP (Xue et al., 2023)	Zero-shot	50.77	49.26	51.12	37.90	47.71	44.75	36.01	45.36
	TDA† (Karmanov et al., 2024)	51.81	51.46	**53.53**	41.65	48.19	43.20	38.04	46.84
	CloudFixer† (Shim et al., 2024)	43.89	41.14	44.75	38.21	35.63	36.32	36.49	39.49
	3DD-TTA† (Dastmalchi et al., 2024)	50.77	47.50	46.47	36.49	45.78	41.31	**42.51**	44.40
	PointCache (Sun et al., 2025)	51.64	50.43	52.67	39.24	48.36	43.89	38.38	46.37
	UniAdapter (Tamjidi et al., 2026)	33.74	34.77	52.50	**42.51**	**49.74**	46.47	34.77	42.07
	LoReTTA	**54.39**	**53.01**	53.36	41.31	**49.74**	**46.64**	35.97	**47.77**
ULIP-2 (Xue et al., 2024)	Zero-shot	44.23	43.20	37.52	31.33	37.87	40.10	21.69	36.56
	TDA† (Karmanov et al., 2024)	**46.99**	**47.50**	41.14	**37.35**	41.65	43.37	25.13	**40.45**
	CloudFixer† (Shim et al., 2024)	29.60	29.95	31.33	21.34	26.33	25.65	26.33	27.22
	3DD-TTA† (Dastmalchi et al., 2024)	36.49	35.28	30.64	22.20	28.92	31.84	**33.22**	31.23
	PointCache (Sun et al., 2025)	45.09	45.09	39.76	32.53	39.24	41.82	23.92	38.21
	UniAdapter (Tamjidi et al., 2026)	42.00	38.04	**45.27**	36.49	**44.41**	**51.64**	23.92	40.25
	LoReTTA	45.44	46.13	39.07	30.12	38.73	39.59	22.72	37.40
Uni3D (Zhou et al., 2024)	Zero-shot	60.24	60.76	57.49	48.54	56.11	53.87	36.83	53.41
	TDA† (Karmanov et al., 2024)	**65.06**	**64.72**	61.45	**54.04**	60.76	58.18	**47.85**	**58.87**
	CloudFixer† (Shim et al., 2024)	49.91	50.43	49.91	38.38	45.96	42.34	46.82	46.25
	3DD-TTA† (Dastmalchi et al., 2024)	53.01	51.64	50.26	38.73	47.33	46.30	45.09	47.48
	PointCache (Sun et al., 2025)	61.79	61.96	58.35	50.77	58.69	54.73	40.31	55.23
	UniAdapter (Tamjidi et al., 2026)	64.89	53.52	**61.62**	49.74	**64.13**	**58.70**	28.74	54.48
	LoReTTA	63.17	62.99	59.72	49.23	59.90	55.39	37.69	55.44

† Adapted baselines evaluated with the same frozen 3D-VLM backbones and evaluation data.

Table 10: **Robustness on ScanObjectNN-C_Obj_BG.** Top-1 accuracy (%) under seven corruptions at severity level 3. Best results are **bolded** and second-best results are underlined.

Backbone	Method	Corruption							Avg.
		Add-G	Add-L	Drop-G	Drop-L	Rot.	Scale	Jitter	
ULIP (Xue et al., 2023)	Zero-shot	45.27	44.41	45.78	35.11	41.14	39.59	31.33	40.38
	TDA† (Karmanov et al., 2024)	46.99	46.82	48.88	38.38	43.89	41.31	32.01	42.61
	CloudFixer† (Shim et al., 2024)	36.83	39.59	38.55	30.29	31.15	31.33	30.64	34.05
	3DD-TTA† (Dastmalchi et al., 2024)	43.72	44.23	43.89	31.84	37.69	37.52	**37.35**	39.46
	PointCache (Sun et al., 2025)	46.64	45.09	46.82	37.18	43.55	40.10	31.33	41.53
	UniAdapter (Tamjidi et al., 2026)	30.64	29.43	**52.32**	**43.72**	40.96	42.34	33.05	38.92
	LoReTTA	**49.05**	**49.05**	50.26	36.32	**44.75**	**43.03**	29.26	**43.10**
ULIP-2 (Xue et al., 2024)	Zero-shot	47.68	47.16	39.41	32.19	40.96	43.37	20.83	38.80
	TDA† (Karmanov et al., 2024)	**51.29**	**51.29**	**44.58**	**36.41**	43.44	45.99	20.86	**41.98**
	CloudFixer† (Shim et al., 2024)	32.36	31.15	29.95	19.62	28.57	27.71	23.41	27.54
	3DD-TTA† (Dastmalchi et al., 2024)	37.18	37.01	27.02	22.72	29.26	32.70	**29.95**	30.83
	PointCache (Sun et al., 2025)	48.36	49.05	42.00	34.60	43.55	45.09	23.24	40.84
	UniAdapter (Tamjidi et al., 2026)	41.82	37.35	43.55	32.19	**52.15**	**46.99**	17.04	38.73
	LoReTTA	49.57	50.60	42.86	28.40	44.06	45.78	20.83	40.30
Uni3D (Zhou et al., 2024)	Zero-shot	61.62	60.76	50.09	43.37	56.45	52.15	27.88	50.33
	TDA† (Karmanov et al., 2024)	**63.72**	**64.89**	53.36	47.57	61.10	56.28	28.60	53.65
	CloudFixer† (Shim et al., 2024)	46.47	45.09	40.45	30.64	42.69	37.18	**38.04**	40.08
	3DD-TTA† (Dastmalchi et al., 2024)	47.68	47.33	38.90	30.12	38.73	39.07	32.87	39.24
	PointCache (Sun et al., 2025)	62.99	62.31	51.98	45.44	57.49	53.87	30.98	52.15
	UniAdapter (Tamjidi et al., 2026)	62.48	53.01	**55.77**	**50.09**	**61.62**	**61.45**	32.19	53.80
	LoReTTA	63.51	63.68	54.91	46.96	60.59	56.42	30.78	**53.84**

† Adapted baselines evaluated with the same frozen 3D-VLM backbones and evaluation data.

A.3 Additional Visualizations

Complementary relation views. Figure 8 visualizes the raw, low-rank, and semantic self-representative matrices under four representative corruptions. The raw relation is more affected by corruption and exhibits noisier cross-class affinities, while the low-rank relation recovers clearer class-wise structure. The semantic view captures complementary category-level dependencies that

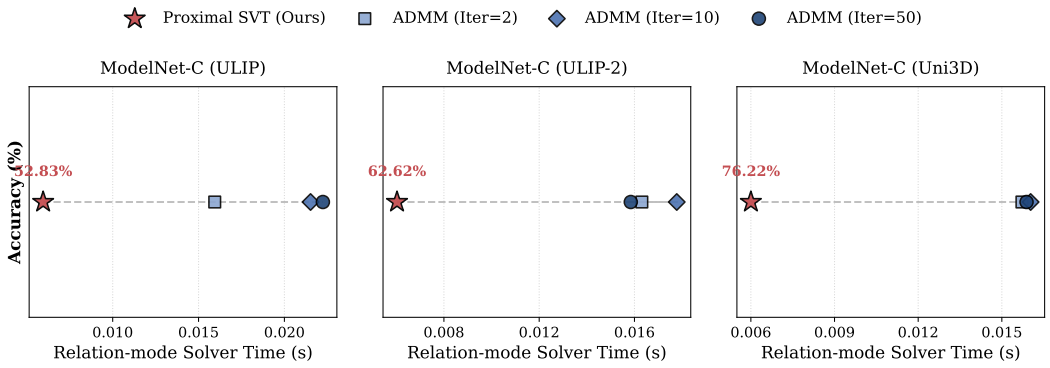

Figure 7: **Efficiency comparison of relation-mode solvers on ModelNet-C.** Proximal SVT and ADMM with different iteration budgets achieve identical downstream accuracy across ULIP, ULIP-2, and Uni3D, while the closed-form proximal update requires substantially less solver time.

Table 11: **Detailed target-set scalability of LoReTTA on ModelNet-C.** Accuracy is reported in %, runtime in seconds, and peak memory in GB. Results for $N <$ Full ($N = 2468$) are averaged over three subset seeds.

	ULIP			ULIP-2			Uni3D		
N	Acc.	Time	Mem.	Acc.	Time	Mem.	Acc.	Time	Mem.
64	41.52 $\pm$ 7.32	0.0079 $\pm$ 0.0013	0.0015	50.60 $\pm$ 7.12	0.0121 $\pm$ 0.0090	0.0015	61.83 $\pm$ 4.35	0.0078 $\pm$ 0.0003	0.0020
128	44.64 $\pm$ 4.24	0.0224 $\pm$ 0.0166	0.0029	54.24 $\pm$ 4.26	0.0129 $\pm$ 0.0060	0.0029	64.51 $\pm$ 2.47	0.0125 $\pm$ 0.0052	0.0036
256	45.98 $\pm$ 2.11	0.0183 $\pm$ 0.0039	0.0080	57.94 $\pm$ 2.23	0.0311 $\pm$ 0.0282	0.0080	70.70 $\pm$ 1.89	0.0257 $\pm$ 0.0091	0.0094
512	49.93 $\pm$ 2.16	0.0566 $\pm$ 0.0162	0.0286	59.43 $\pm$ 1.50	0.0806 $\pm$ 0.0106	0.0286	73.08 $\pm$ 0.92	0.0903 $\pm$ 0.0847	0.0315
1024	51.60 $\pm$ 0.68	0.0616 $\pm$ 0.0230	0.1050	60.00 $\pm$ 0.38	0.0811 $\pm$ 0.0225	0.1050	75.47 $\pm$ 0.82	0.1135 $\pm$ 0.0253	0.1138
Full	**52.83**	0.0825	0.5945	**62.62**	0.1083	0.5945	**76.22**	0.1259	0.6123

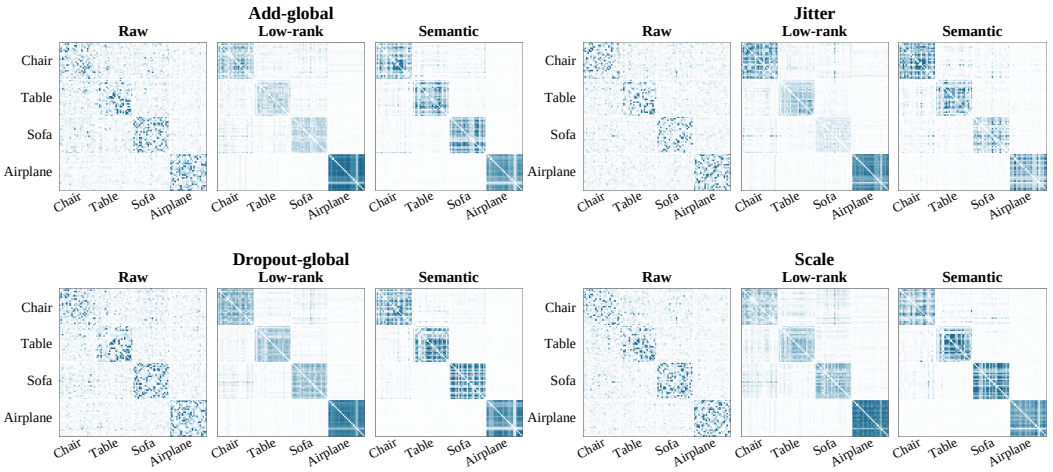

Figure 8: **Complementary relation views.** Raw, low-rank, and semantic self-representation matrices under representative ModelNet-C corruptions.

are not fully reflected by the structural relation. The consistent patterns across different corruptions support combining multiple relation views for robust target-domain modeling.