ROBUST TEST-TIME TEXT–3D SCENE RETRIEVAL: BENCHMARK AND RELIABILITY-ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployed Text–3D Scene Retrieval (T3SR) systems receive unlabeled queries that deviate from the training distribution: typos and paraphrases on the text side, and occlusion, sensor noise, and partial scans on the 3D side. Test-time adaptation (TTA) can help address these distribution shifts, but its effectiveness depends on the reliability of self-generated cross-modal correspondences. Two coupled challenges arise: gallery-level matching biases can undermine correspondence reliability despite high prediction confidence, while shared query-encoder updates can induce representation drift in uncertain queries even when they are excluded from the adaptation objective. To systematically study these challenges, we introduce M3SP, a benchmark comprising 14 multi-level 3D perturbation types, each instantiated at five severity levels across four datasets. We further propose **STAR-3D** (**S**tructure-guided **T**est-time **A**daptation for Text–3D Scene **R**etrieval). STAR-3D first calibrates query–gallery matching scores through history-aware Sinkhorn normalization, then assesses provisional correspondences by comparing query and candidate representations in a centered residual space. Structural consistency and a memory-adaptive confidence gate jointly determine which queries drive confidence sharpening, while the remaining queries are regularized toward their frozen-source representations. Extensive experiments across datasets demonstrate that STAR-3D outperforms competitive TTA baselines overall under query shift and query–gallery shift. Code is available at `https://anonymous.4open.science/r/STAR-3D-C486`.

1 INTRODUCTION

Text–3D Scene Retrieval (T3SR) matches natural-language descriptions to 3D point-cloud scenes and retrieves descriptions for observed scenes (Feng et al., 2025c; Ye et al., 2025). In deployment, typos and paraphrases can change text queries, while sensor noise, occlusion, and incomplete scans can change 3D queries. These shifts can alter retrieval rankings even when the intended correspondence is unchanged. Target correspondence labels are unavailable, and the gallery encoder and indexed embeddings remain fixed during each target stream; online adaptation of the query-side model must therefore rely on its own predictions.

Cross-modal test-time adaptation (TTA) offers a way to respond to query shift. Li et al. (2025) study online adaptation for image–text retrieval with TCR, using refined predictions and embedding adjustment; Zhang et al. (2026) study video–text retrieval with HAT-VTR, including the concentration of matches on a few gallery items under perturbation. For scene-level T3SR, point measurement errors, missing object evidence, and scene-wide acquisition failures also need a controlled evaluation protocol. Beyond this evaluation gap, learning from self-generated correspondences raises a shared problem across modalities: incorrect matches can be reinforced, while an encoder update can affect queries that supplied no adaptation signal.

This problem involves two linked decisions. (1) **Which correspondences should guide learning?** Under paired clean and corrupted text queries, the same 20 gallery scenes receive 8.3% and 18.3% of Top-1 assignments, respectively (Fig. 1a,b). This concentration motivates correcting gallery-level matching preferences, but it does not establish whether any individual match is correct. Among the most confident 20% of raw predictions in our ScanRefer diagnostics, 62.3% of *t2s* and 43.4% of *s2t* Top-1 pairs are incorrect (Fig. 1c), motivating a separate assessment of query–candidate consis-

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.

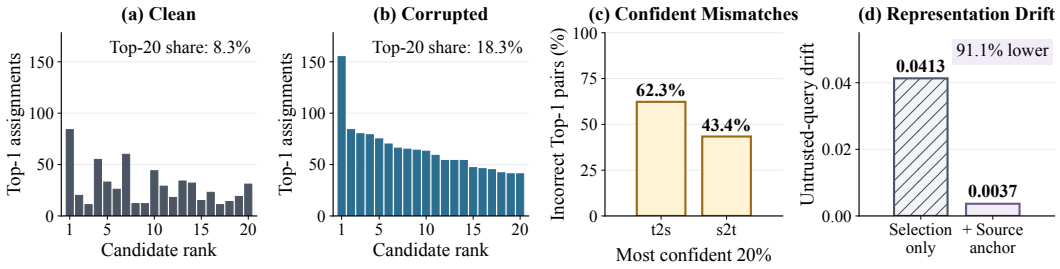


Figure 1: **Challenges in adapting cross-modal retrieval. (a,b)** Clean and corrupted text queries, respectively, evaluated with the same frozen source model reveal increased concentration of Top-1 assignments. Both panels show the same 20 scene candidates, ranked by corrupted-query assignment counts, with identical axis ranges; annotations report their shares of all assignments in each condition. **(c)** At the correspondence level, even the lowest-entropy 20% of raw predictions contain substantial errors, motivating reliability assessment beyond prediction confidence. **(d)** Restricting adaptation to selected queries still leaves the source-defined untrusted group susceptible to representation drift; frozen-source anchoring reduces this cosine drift, motivating explicit preservation alongside selective adaptation.

tency. (2) **What happens to queries without reliable supervision?** Excluding a query from the adaptation loss does not exclude it from the effects of adaptation: updates to shared query-encoder parameters can still move its representation. As shown in Figure 1(d), this drift for a fixed, source-defined untrusted group and shows that a frozen-source anchor reduces it. The resulting question is both which correspondences should drive learning and how to constrain queries for which reliable supervision is unavailable.

To study these shifts in T3SR, we introduce **M3SP** (**M**ulti-level **3**D **S**cene **P**erturbations). It provides 14 perturbation types at five severity levels across ScanRefer (Chen et al., 2020), SceneDepict (Feng et al., 2025c), Nr3D, and Sr3D (Achlioptas et al., 2020). Its micro-, meso-, and macro-level conditions perturb point measurements, coherent object regions, and scene-wide acquisition or geometry–appearance relations, respectively. This common retrieval protocol makes it possible to compare how models respond to different forms of 3D evidence loss.

We propose **STAR-3D** (**S**tructure-guided **T**est-time **A**daptation for Text–3D Scene **R**etrieval) to connect three decisions: calibrate, route, and adapt/preserve. First, following prior work on matching calibration (Zhang et al., 2026; Pan et al., 2025), recent query statistics support Sinkhorn calibration of gallery-level matching preferences before provisional correspondences are formed. Next, centered query and candidate embeddings provide a cross-modal consistency signal, which is combined with prediction confidence to route queries. Here, structural consistency refers to relative geometry in the cross-modal embedding space, rather than the point–object–scene hierarchy used to organize M3SP. Admitted queries drive confidence sharpening; the remaining queries are regularized toward their frozen-source representations instead of uncertain gallery matches. Thus, every query is assigned an adaptation or preservation objective according to the available evidence.

Under level-5 3D query corruptions with CrossOver, the complete STAR-3D pipeline achieves R@1 of 23.55% in *t2s* and 54.54% in *s2t* on ScanRefer, outperforming the strongest comparison methods by 1.28 and 9.09 percentage points, respectively. We also evaluate Query Shift (QS) and Query-Gallery Shift (QGS), including cross-dataset and zero-shot settings, and report component diagnostics and ablations separately.

Our contributions are summarized as follows:

- We introduce M3SP, a benchmark with 14 perturbation types at five severity levels across four datasets. Its point-, object-, and scene-level conditions test distinct forms of degraded 3D evidence within a common retrieval protocol.
- We propose STAR-3D, which calibrates provisional matches, routes queries using centered cross-modal structure and confidence, and couples adaptation of admitted queries with frozen-source anchoring of the remainder.
- We evaluate the complete retrieval pipeline across datasets and directions under QS and QGS, and report ablations and diagnostics for calibration, routing, and representation drift.

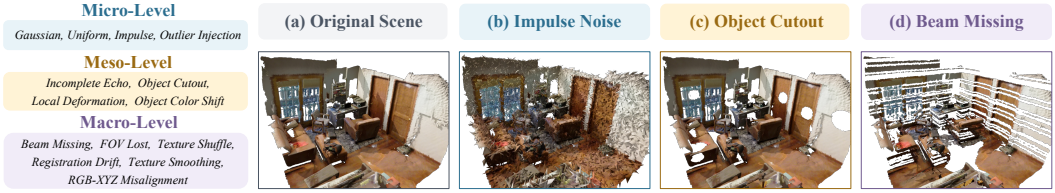


Figure 2: Overview of the M3SP benchmark across Micro-, Meso-, and Macro-Level perturbations. Representative examples are shown, with full benchmark details provided in the Appendix B.

2 Related Work

Text–3D Scene Retrieval. Text–3D scene retrieval (T3SR) aims to connect natural-language descriptions with 3D environments. Initial works focus on global cross-modal alignment in a shared space (Yu et al., 2024; Feng et al., 2025c; Sarkar et al., 2025). To handle complex spatial relations, subsequent methods introduce structured representations such as scene graphs (Chen et al., 2024). Recent studies further address partial correspondences via submap matching (Feng et al., 2025a) or interactive query expansion (Feng et al., 2025b). Despite this progression, existing methods assume clean, stationary test distributions and keep models fixed at inference. In practice, geometric, density, and appearance corruptions can severely disrupt pretrained Text–3D alignment. To tackle this overlooked issue, we investigate T3SR under test-time distribution shifts and propose an adaptation framework that continuously recalibrates cross-modal alignment using only unlabeled test queries.

Test-Time Adaptation. Test-time adaptation (TTA) updates pre-trained models on unlabeled test samples at inference without target labels (Sun et al., 2020; Wang et al., 2020). Early TTA mainly targets classification via entropy minimization (Wang et al., 2020), sample selection (Niu et al., 2022), and consistency regularization (Wang et al., 2022; Niu et al., 2023). Recent efforts extend TTA to instance-level retrieval, including image retrieval (Ma et al., 2024), cross-modal query shift (Li et al., 2025), severe corruption (Yang et al., 2026), and video-text ranking (Zhang et al., 2026). These works demonstrate that retrieval TTA must preserve the pre-trained embedding space structure during online adaptation, rather than naively minimizing prediction uncertainty on shifted queries. However, existing cross-modal TTA focuses on 2D image and video domains, leaving T3SR underexplored. 3D queries exhibit distinct shifts in geometry, density, and visibility, while requiring structured spatial understanding. Moreover, current 3D TTA methods target recognition (Wei et al., 2025) or segmentation (Shin et al., 2022), lacking support for cross-modal ranking. These disparities motivate a specialized TTA framework for robust T3SR.

3 M3SP: Multi-level 3D Scene Perturbation Benchmark

To systematically investigate the vulnerabilities of T3SR models under realistic 3D query shifts and establish criteria for evaluating adaptation methods, it is critical to construct a benchmark tailored to the inherent properties of point-cloud modalities. We introduce M3SP benchmark, which defines 14 corruption types at five calibrated severity levels, yielding 70 controlled conditions. These conditions are instantiated on the evaluation splits of four ScanNet-based retrieval datasets: ScanRefer (Chen et al., 2020), SceneDepict (Feng et al., 2025c), Nr3D and Sr3D (Achlioptas et al., 2020). We organize them by the representation dependency they disrupt: individual measurements at the micro level, coherent object instances at the meso level, and scene-level sampling or cross-attribute spatial relations at the macro level.

Micro-level perturbations alter individual point measurements while preserving object identity and the overall scene layout. They simulate low-level failures of depth sensors and early signal processing: electronic and ranging uncertainty produces dense coordinate fluctuations, whereas multipath reflection, cross-sensor interference, or erroneous depth estimates introduce sparse abnormal returns. Accordingly, we adopt *Gaussian* and *Uniform Noise* to model stochastic and bounded ranging errors, *Impulse Noise* for sporadic readout failures, and *Outlier Injection* for spurious measurements. These corruptions modify XYZ without using semantic instances, thereby isolating robustness to sensor-level noise from changes in object or scene structure.

Meso-level perturbations target coherent object regions, where degradation can arise during both acquisition and reconstruction. Dark or weakly reflective surfaces, self-occlusion, and restricted viewpoints can suppress depth returns and leave objects partially observed; imperfect multi-view fusion can warp local geometry, while illumination, white balance, and RGB calibration alter object appearance. Guided by ScanNet instance masks, we model these failures with *Incomplete Echo* for dispersed missing returns, *Object Cutout* for contiguous depth holes, *Local Deformation* for reconstruction-induced shape errors, and *Object Color Shift* for photometric variation. This instance-aware construction concentrates each degradation on a coherent semantic entity, providing a more realistic test of object-level robustness than indiscriminate scene-wide corruption.

Macro-level perturbations disrupt scene-wide acquisition patterns or the higher-order correspondence between geometry and appearance. Such failures arise when sensor obstruction or hardware dropout removes entire scan regions, when a restricted field of view truncates the observed scene, or when pose estimation, texture processing, and cross-sensor calibration introduce spatially coherent errors. We model these factors with *Beam Missing* and *FOV Lost* for structured data absence, *Registration Drift* for accumulated multi-view alignment error, and *Texture Smoothing*, *Texture Shuffle*, and *RGB–XYZ Misalignment* for degraded or misplaced appearance. Unlike local point noise or isolated object corruption, these perturbations probe global scene organization and the spatial binding required for reliable cross-modal retrieval.

4 METHODOLOGY

In this section, we introduce STAR-3D for test-time adaptation in T3SR (Figure 3). Section 4.1 defines the retrieval setting and presents entropy minimization as a basic adaptation objective. Section 4.2 explains how recent query statistics adjust gallery-level matching preferences. Section 4.3 then selects provisional pairs and evaluates their centered-residual consistency to identify structurally eligible queries. Section 4.4 combines confidence-based admission with complementary sharpening and frozen-source anchoring objectives.

4.1 PROBLEM FORMULATION

In T3SR, a query $x^Q \in \mathcal{X}^Q$ retrieves its matching item from a gallery $\mathcal{X}^G = \{x_j^G\}_{j=1}^{N_G}$. The scene-to-text (*s2t*) and text-to-scene (*t2s*) directions interchange the query and gallery modalities. A source dual encoder $f_{\Theta_s} = (f^Q_{\Theta_s}, f^G_{\Theta_s})$ maps both modalities into a shared d-dimensional embedding space. Hereafter, t indexes incoming query batches, and i indexes queries within each batch. Let $\mathbf{H}_t^Q \in \mathbb{R}^{B\times d}$ and $\mathbf{H}^G \in \mathbb{R}^{N_G\times d}$ collect the ℓ_2-normalized embeddings of the current B queries and the N_G gallery items, respectively. Gallery embeddings $\mathbf{H}^G$ are precomputed and held fixed throughout adaptation. The similarity matrix is $\mathcal{S}_t = \mathbf{H}_t^Q(\mathbf{H}^G)^\top$, yielding full-gallery query predictions

$$\mathbf{p}_t = \mathrm{Softmax}(\mathcal{S}_t/\tau)\,, \tag{1}$$

where Softmax is applied row-wise and τ is the prediction temperature. Standard entropy-based TTA updates a parameter subset $\theta \subseteq \Theta_s$, restricted to the query encoder, by minimizing

$$\mathcal{L}_{\mathrm{TTA}}(\theta) = \eta(\mathbf{p}_t(\theta))\,, \tag{2}$$

where $\eta(\mathbf{p}_t) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{N_G} p_{t,ij}\log p_{t,ij}$ denotes the mean prediction entropy.

Following prior work (Li et al., 2025), we consider two distribution-shift scenarios. Let $\mathcal{P}_S$, $\mathcal{P}_F$, and $\mathcal{P}_T$ denote the source, fine-tuning, and target data distributions, respectively, with superscripts Q and G indicating their query and gallery marginals. **Query Shift (QS)** affects only the query distribution, i.e., $\mathcal{P}_T^Q \neq \mathcal{P}_S^Q$ and $\mathcal{P}_T^G = \mathcal{P}_S^G$, motivating query-encoder adaptation. The more challenging **Query-Gallery Shift (QGS)** affects both query and gallery distributions and includes two cases: (1) *zero-shot adaptation*, which directly adapts the source-pretrained model to the target domain under $\mathcal{P}_T \neq \mathcal{P}_S$; and (2) *cross-dataset adaptation*, which adapts a model fine-tuned on dataset $\mathcal{D}_F$ to the target domain under $\mathcal{P}_T \neq \mathcal{P}_F$.

4.2 ONLINE MATCHING SCORE CALIBRATION

Figure 1(a,b) shows that text query corruption increases the concentration of top-1 assignments on a small set of gallery scenes. Such concentration can bias candidate selection based on raw

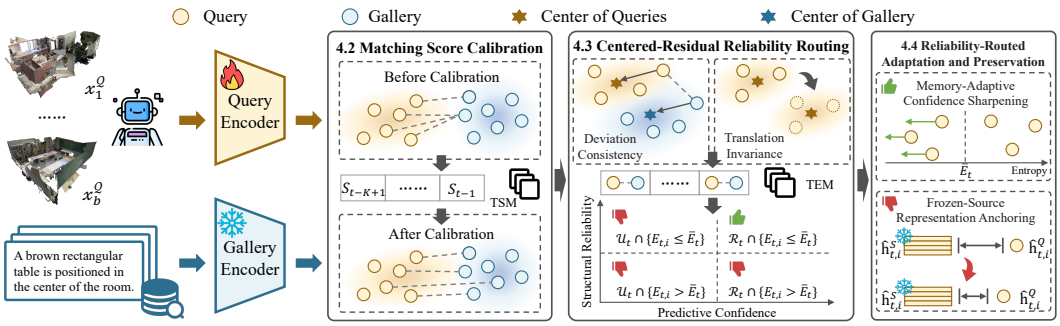


Figure 3: Overview of STAR-3D, which performs matching score calibration and guides online adaptation through reliability-aware routing, sharpening reliable predictions while anchoring uncertain query representations to the frozen source encoder.

similarities (Lin et al., 2025); we therefore calibrate the score matrix $\mathcal{S}_t$ used in Eq. (1) before forming provisional correspondences. Following prior work on online retrieval calibration (Zhang et al., 2026), we maintain a Temporal Similarity Memory (TSM) $\mathcal{M}^S_{t-1}$ that retains up to $K-1$ recent batches of finalized scores, where K denotes the calibration window size. Concatenating the stored scores with the current scores yields

$$\bar{\mathcal{S}}_t = \mathrm{concat}\big(\mathcal{M}^S_{t-1}, \mathcal{S}_t\big)\,. \tag{3}$$

Applying Sinkhorn matrix scaling (Cuturi, 2013; Pan et al., 2025) yields

$$\mathbf{\Pi}^\star_t = \mathrm{diag}(\boldsymbol{\alpha}_t)\exp\big(\bar{\mathcal{S}}_t/\tau_s\big)\,\mathrm{diag}(\boldsymbol{\beta}_t), \tag{4}$$

where τ_s is the calibration temperature, exp is applied element-wise, and the scaling vectors are obtained by alternating row and column normalization. Equivalently,

$$\begin{aligned} \tau_s \log \Pi^\star_{t,ij} &= \bar{\mathcal{S}}_{t,ij} + \hbar^Q_{t,i} + \hbar^G_{t,j}, \\ \hbar^Q_{t,i} &= \tau_s \log \alpha_{t,i}, \qquad \hbar^G_{t,j} = \tau_s \log \beta_{t,j}. \end{aligned} \tag{5}$$

Within each row, the query term is constant across gallery items, whereas the gallery term adjusts their relative scores based on recent query statistics. We denote the current-batch rows of $\mathbf{\Pi}^\star_t$ by $\hat{\mathcal{S}}_t$. When the TSM is empty, we set $\hat{\mathcal{S}}_t = \mathcal{S}_t$. These scores define provisional candidates; a separate consistency check determines which pairs are suitable for learning.

4.3 Centered-Residual Reliability Routing

For each query, we select the gallery candidate with the highest calibrated score:

$$j^\star_{t,i} = \arg\max_j \hat{\mathcal{S}}_{t,ij}, \qquad \mathbf{c}_{t,i} = \mathbf{h}^G_{j^\star_{t,i}}. \tag{6}$$

These pairs remain provisional: a concentrated or high-confidence prediction can still be incorrect. The raw query–candidate distance also reflects the offset between modality centers (Wang et al., 2026). Let $\bar{\mathbf{h}}^Q_t = \frac{1}{B}\sum_{i=1}^B \mathbf{h}^Q_{t,i}$ and $\bar{\mathbf{c}}_t = \frac{1}{B}\sum_{i=1}^B \mathbf{c}_{t,i}$ denote the query and selected-candidate centers, respectively. With centered residuals $\mathbf{r}^Q_{t,i} = \mathbf{h}^Q_{t,i} - \bar{\mathbf{h}}^Q_t$ and $\mathbf{r}^G_{t,i} = \mathbf{c}_{t,i} - \bar{\mathbf{c}}_t$, we define the Centered Residual Structural Inconsistency (CRSI) score as

$$\delta^{\mathrm{CRSI}}_{t,i} = 2\big\|\mathbf{r}^Q_{t,i} - \mathbf{r}^G_{t,i}\big\|_2 - \big\|\mathbf{r}^Q_{t,i}\big\|_2 - \big\|\mathbf{r}^G_{t,i}\big\|_2. \tag{7}$$

CRSI evaluates pairwise discrepancy using modality-centered residuals. For fixed candidate assignments, the score is invariant to uniform translations within either modality. By balancing residual mismatch against each embedding's distance from its modality center, it prioritizes provisional pairs with consistent relative configurations across modalities.

We select the m lowest-scoring pairs and partition the query indices as

$$\begin{aligned} \mathcal{R}_t &= \mathrm{Bottom}_m\big(\{\delta^{\mathrm{CRSI}}_{t,i}\}_{i=1}^B\big)\,, \\ m &= \max(1, \lfloor \rho B \rfloor)\,, \qquad \mathcal{U}_t = \{1,\dots,B\} \setminus \mathcal{R}_t, \end{aligned} \tag{8}$$

where ρ is the reliable-sample ratio and Bottom_m returns the indices of the m smallest scores. This partition determines structural eligibility: queries in $\mathcal{R}_t$ must pass an additional confidence check before contributing to learning, whereas those in $\mathcal{U}_t$ are subject to a preservation constraint.

4.4 Reliability-Routed Adaptation and Preservation

As illustrated in Fig. 1(d), updates driven by selected queries can also alter the representations of excluded queries through the shared encoder. We therefore assess prediction confidence within $\mathcal{R}_t$ and assign each query either a confidence-sharpening or a representation-preservation objective.

Memory-Adaptive Confidence Sharpening. Let $\mathbf{C}_t = [\mathbf{c}_{t,1}, \ldots, \mathbf{c}_{t,B}]^\top \in \mathbb{R}^{B\times d}$ denote the matrix formed by stacking the selected candidate embeddings. For each query, we compute a distribution over the selected candidates and assess confidence using its entropy:

$$\widehat{\mathbf{p}}_{t,i} = \mathrm{Softmax}\left(\frac{\mathbf{C}_t \mathbf{h}^Q_{t,i}}{\tau}\right), \qquad E_{t,i} = -\sum_{j=1}^{B} \widehat{p}_{t,ij} \log \widehat{p}_{t,ij}. \tag{9}$$

Before optimization, we append the detached pre-adaptation entropies of queries in $\mathcal{R}_t$ to a bounded Temporal Entropy Memory (TEM) $\mathcal{M}^E_t$. The memory determines the threshold

$$\bar{E}_t = \max_{E\in\mathcal{M}^E_t} E, \tag{10}$$

which remains fixed throughout the batch. At each optimization step, we define an admission indicator, a detached confidence weight, and the weighted entropy loss:

$$\begin{aligned} I_{t,i} &= \mathbb{I}\big[i \in \mathcal{R}_t \wedge E_{t,i} \le \bar{E}_t\big], \qquad W_{t,i} = \left[1 - \frac{\mathrm{sg}(E_{t,i})}{\bar{E}_t + \epsilon}\right]_+, \\ \mathcal{L}_{\mathrm{MACS}} &= \frac{1}{\max(1, \sum_{i=1}^{B} I_{t,i})} \sum_{i=1}^{B} I_{t,i}\, W_{t,i}\, E_{t,i}, \end{aligned} \tag{11}$$

where $[z]_+ = \max(z, 0)$, sg denotes the stop-gradient operator, and $\epsilon > 0$ ensures numerical stability. Structurally selected queries with sufficiently low entropy thus drive confidence-weighted entropy minimization.

Frozen-Source Representation Anchoring. Let $\mathcal{A}_t = \{i : I_{t,i} = 1\}$ denote the admitted query indices. All remaining queries are assigned to the preservation set:

$$\mathcal{P}_t = \{1, \ldots, B\} \setminus \mathcal{A}_t = \mathcal{U}_t \cup \{i \in \mathcal{R}_t : E_{t,i} > \bar{E}_t\}. \tag{12}$$

For each target query $x^Q_{t,i}$, let $\mathbf{h}^Q_{t,i} = f^Q_{\Theta_t}(x^Q_{t,i})$ and $\mathbf{h}^S_{t,i} = f^Q_{\Theta_s}(x^Q_{t,i})$ denote its online and frozen-source features, respectively. With $\widehat{\mathbf{h}} = \frac{\mathbf{h}}{\|\mathbf{h}\|_2}$, we define the preservation loss:

$$\mathcal{L}_{\mathrm{RA}} = \frac{1}{\max(1, |\mathcal{P}_t|)} \sum_{i\in\mathcal{P}_t} \left[1 - (\widehat{\mathbf{h}}^Q_{t,i})^\top \mathrm{sg}(\widehat{\mathbf{h}}^S_{t,i})\right]. \tag{13}$$

This anchor does not rely on gallery assignments and constrains representation drift in queries that do not pass the admission criteria. Within each batch, the candidates, CRSI partition, and threshold $\bar{E}_t$ remain fixed, whereas the entropies $E_{t,i}$ and the resulting objective assignments are recomputed at each optimization step. Since the pre-adaptation entropies of the current structurally selected queries enter TEM before the first step, the confidence mechanism generally begins by reweighting these queries rather than further filtering them.

Overall Objective. The routed objective extends the entropy-based formulation in Eq. (2):

$$\mathcal{L}_{\mathrm{TTA}} = \mathcal{L}_{\mathrm{MACS}} + \mathcal{L}_{\mathrm{RA}}, \tag{14}$$

After updating the query-side parameters, we store the batch's detached post-adaptation raw scores in TSM and carry the updated encoder to the next batch. When enabled for evaluation, final Sinkhorn calibration is applied separately to the completed score matrix without affecting online routing or parameter updates. Numerical and queue-management details appear in Appendix A.5.

Table 1: s2t results (R@1, %) on ScanRefer under level-5 corruption severity across various baseline methods. The best results are shown in **bold**, and the second-best results are underlined.

	Micro-Level				Meso-Level				Macro-Level						
Method	GN	OI	UN	IN	IE	OC	LD	OCS	BM	FL	RD	TSm	TSh	RXM	**Avg.**
CrossOver	46.51	30.16	38.69	20.77	43.24	46.23	45.80	40.97	30.44	28.59	45.80	45.55	45.00	45.52	39.52
+ Tent	46.66	30.44	38.12	20.75	43.53	46.09	45.75	41.44	30.58	28.88	45.95	45.24	45.10	45.38	39.56
+ READ	46.80	30.27	38.15	20.63	43.95	45.80	45.13	40.83	30.73	29.02	46.09	45.24	45.23	45.38	39.51
+ SAR	46.51	30.17	38.69	20.88	43.24	46.23	45.80	41.06	30.30	28.59	45.94	45.64	45.77	45.66	39.61
+ EATA	46.51	31.15	39.40	20.91	43.81	45.38	45.09	41.59	30.44	29.02	46.09	45.10	45.24	45.51	39.66
+ TCR	50.36	35.99	42.25	34.14	46.94	45.23	44.95	40.83	34.99	33.14	48.79	45.80	45.94	45.80	42.51
+ HAT-VTR	50.64	40.83	48.93	37.98	50.92	44.67	49.36	38.69	35.28	38.12	49.93	50.36	50.36	50.21	45.45
+ Ours	**57.47**	**48.65**	**57.75**	**53.49**	**57.61**	**58.32**	**57.47**	**50.50**	**46.37**	**44.24**	**57.75**	**58.46**	**58.61**	**56.90**	**54.54**
Mosaic3D	15.37	20.77	7.25	16.36	20.63	20.91	20.91	12.66	13.09	18.92	21.48	21.34	3.41	20.91	16.66
+ Tent	15.37	20.63	7.25	16.08	20.48	20.91	20.63	12.52	12.95	18.78	21.19	20.91	3.70	20.34	16.50
+ READ	15.65	20.77	7.39	16.79	20.91	20.77	21.05	12.66	13.23	18.92	21.76	21.48	3.41	20.91	16.72
+ SAR	14.94	20.34	7.11	15.79	20.20	20.06	20.06	12.23	12.66	18.21	20.77	20.63	3.41	20.20	16.07
+ EATA	14.65	19.92	6.97	15.36	19.20	20.06	20.20	12.38	12.23	17.92	20.06	20.20	3.13	19.77	15.77
+ TCR	15.65	21.05	7.82	16.79	21.19	21.76	22.05	13.51	14.94	19.35	21.48	22.05	4.13	21.48	17.37
+ HAT-VTR	19.77	25.75	9.67	23.47	26.46	26.74	28.17	20.77	16.07	23.04	**26.88**	**28.17**	8.96	27.31	22.27
+ Ours	**21.34**	**27.60**	**13.09**	**24.04**	**28.45**	**27.45**	**29.16**	**22.48**	**19.06**	**23.90**	26.74	28.02	**11.10**	**27.88**	**23.55**

Table 2: Results on t2s (R@1, %) on the ScanRefer under text perturbations.

	Character-Level					Word-Level					Sentence-Level					
Method	OCR	CI	CR	CS	CD	SR	RI	RS	RD	IP	Formal	Casual	Passive	Active	Backtrans	**Avg.**
CrossOver	12.92	6.51	8.01	10.33	10.47	15.76	17.80	22.13	21.56	26.17	23.19	20.26	21.34	27.16	26.61	18.01
+ Tent	10.13	3.91	5.08	7.30	8.38	14.21	13.68	18.99	19.02	21.25	20.57	16.93	18.83	23.60	23.87	15.05
+ READ	14.13	7.48	8.78	11.01	11.83	17.07	19.10	22.92	22.76	25.66	23.80	20.26	21.48	27.33	28.12	18.78
+ SAR	11.78	5.80	7.50	9.17	9.91	15.14	17.89	21.02	20.83	25.04	22.57	19.32	20.91	26.57	26.17	17.31
+ EATA	9.82	3.67	4.78	6.59	7.16	13.51	14.88	16.09	17.70	22.45	21.12	14.67	15.41	24.17	21.75	14.25
+ TCR	14.85	8.05	**9.53**	11.64	12.42	17.64	20.06	23.63	24.08	27.97	24.54	21.37	22.86	29.67	28.53	19.79
+ HAT-VTR	14.41	**8.73**	9.50	11.27	11.74	17.20	18.89	23.09	22.99	26.81	22.92	20.34	22.35	28.61	27.81	19.11
+ Ours	**15.29**	8.15	9.27	**12.20**	**12.87**	**19.57**	**22.08**	**25.75**	**26.16**	**30.36**	**26.51**	**23.29**	**24.25**	**31.88**	**31.34**	**21.26**
Mosaic3D	9.74	8.95	8.38	9.33	8.98	11.07	11.71	11.72	11.64	12.23	11.62	11.93	11.14	11.82	12.28	10.84
+ Tent	9.19	8.11	7.20	8.49	8.41	10.40	11.34	11.39	11.05	11.75	11.29	11.72	10.81	11.38	12.05	10.31
+ READ	9.83	8.96	8.38	8.90	9.05	10.98	11.61	11.89	11.44	11.99	11.54	11.91	10.95	11.59	11.98	10.73
+ SAR	7.71	6.63	6.43	7.18	7.01	7.67	9.36	7.94	8.88	9.69	8.75	10.53	8.78	8.69	10.40	8.38
+ EATA	7.97	7.13	6.22	7.51	6.81	9.27	10.24	10.16	10.57	11.04	11.22	11.18	9.74	11.00	11.74	9.45
+ TCR	9.93	9.17	8.32	9.12	9.00	10.91	11.89	11.71	11.54	12.26	11.58	11.89	11.10	11.86	12.43	10.85
+ HAT-VTR	9.50	8.26	8.26	8.72	8.85	10.33	11.07	10.83	10.74	11.37	10.84	11.29	10.16	10.65	11.76	10.18
+ Ours	**11.14**	**9.89**	**9.33**	**10.14**	**9.82**	**11.71**	**12.80**	**12.82**	**12.19**	**13.30**	**12.73**	**12.86**	**11.93**	**12.83**	**13.51**	**11.80**

5 EXPERIMENTS

5.1 EXPERIMENT SETTINGS

Base Retrieval Models, TTA Baselines, and Benchmark. We use CrossOver (Sarkar et al., 2025) and Mosaic3D (Lee et al., 2025) as the base T3SR models. For comparison, we consider six competitive TTA methods, which can be grouped into two categories: (1) general-purpose TTA methods, including Tent (Wang et al., 2020), EATA (Niu et al., 2022), SAR (Niu et al., 2023), and READ (**?**); and (2) TTA methods specifically designed for cross-modal retrieval, namely TCR (Li et al., 2025) and HAT-VTR (Zhang et al., 2026). All methods are evaluated on our proposed M3SP benchmark under the same evaluation protocol.

Evaluation Protocol. We evaluate both scene-to-text (s2t) and text-to-scene (t2s) retrieval under Query Shift (QS) and Query-Gallery Shift (QGS) (encompassing both *cross-dataset* and *zero-shot* settings). For s2t retrieval, scene corruptions are adopted from our M3SP benchmark. For t2s retrieval, we construct character-, word-, and sentence-level text perturbations following TCR Li et al. (2025). Performance is reported in Recall@K, averaged uniformly across all corruption conditions.

Implementation Details. For both base T3SR models, we freeze all model components and update only the affine parameters of the normalization layers in the query encoder. Specifically, for t2s retrieval, we adapt the Layer Normalization (LN) affine parameters of the text encoder, whereas for s2t retrieval, we adapt the Batch Normalization (BN) affine parameters of the 3D encoder while keeping its running statistics fixed. Online adaptation is performed using AdamW with a batch size of 32, a learning rate of 10^{-3}. We set $\tau = 0.1$, $\tau_s = 0.05$, $\rho = 0.3$, and $K = 4$, with a 32-entry

TEM. All experiments run on a single NVIDIA RTX 4090 GPU, with the model reset to its initial checkpoint for each corruption.

5.2 Comparison Results on Query Shift

We first evaluate STAR-3D under QS, where only query instances are corrupted at test time while the gallery remains clean. This setting reflects common real-world scenarios, such as retrieving from a clean gallery using a noisy 3D scene scan or a text query containing typographical errors. As shown in Tables 1, the retrieval performance of the base models degrades substantially under the multi-level 3D perturbations in M3SP. Conventional TTA methods, such as Tent, EATA, SAR, and READ, provide only marginal improvements, highlighting the difficulty of adapting cross-modal alignment in 3D scene-text retrieval. Retrieval-specific TTA methods such as TCR achieve partial recovery, whereas STAR-3D consistently delivers superior robustness across diverse perturbation types and outperforms all competing methods on s2t retrieval. For the reverse t2s task, as reported in Table 2, STAR-3D similarly achieves the best overall performance under text-side query corruptions. These results demonstrate the effectiveness of our adaptation strategy under query distribution shifts. Due to space limitations, detailed per-perturbation results and evaluations on additional datasets are provided in the Appendix C.

Table 3: *Cross-dataset Adaptation* Performance.

	ScanRefer → SceneDepict		**SceneDepict → ScanRefer**	
Method	*s2t* R@1	*t2s* R@1	*s2t* R@1	*t2s* R@1
CrossOver	8.23	4.75	11.32	5.95
+ Tent	7.97	2.57	11.38	5.83
+ READ	8.04	4.03	11.43	5.91
+ SAR	8.21	2.66	11.34	5.90
+ EATA	8.19	2.19	11.51	5.28
+ TCR	8.74	4.26	11.97	5.86
+ HAT-VTR	9.36	5.28	12.46	5.35
+ Ours	**9.73**	**6.13**	**13.09**	**6.50**

5.3 Comparison Results on Query–Gallery Shift

We further evaluate STAR-3D under the two QGS settings. In cross-dataset adaptation (Table 3), STAR-3D achieves the highest R@1 in both retrieval directions for both transfer pairs. In zero-shot adaptation (Table 4), absolute R@1 remains low across methods. This is consistent with the difficulty of transferring pretrained representations without task-specific supervised alignment: broad semantic similarity does not necessarily identify the exact scene–text correspondence required by R@1, particularly when both modalities differ from the pretraining distribution. Nevertheless, STAR-3D achieves the best performance across all six evaluations, demonstrating the effectiveness of the overall framework under QGS. These results support the design motivation in our method: calibrating gallery-level matching bias, routing adaptation through structurally reliable and confident correspondences, and preserving uncertain queries through frozen-source anchoring, rather than indiscriminately sharpening potentially incorrect predictions.

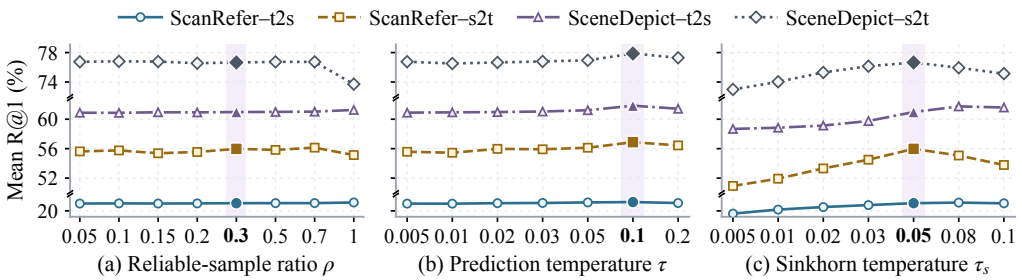
Table 4: Results of *Zero-shot Adaptation*.

	ScanRefer		**SceneDepict**	
Method	*s2t* R@1	*t2s* R@1	*s2t* R@1	*t2s* R@1
Mosaic3D	3.98	2.29	6.63	3.07
+ Tent	3.23	0.70	5.33	1.01
+ READ	4.43	2.25	6.66	4.53
+ SAR	3.87	1.96	6.63	3.07
+ EATA	3.87	1.96	6.63	3.07
+ TCR	0.38	1.19	0.89	2.20
+ HAT-VTR	1.15	2.09	3.23	4.07
+ Ours	**6.40**	**3.83**	**6.97**	**5.12**

5.4 Ablation and Analytic Study

Sensitivity Analysis. Figure 4 examines the effects of three hyperparameters through separate sweeps. Performance is relatively stable for the reliable-sample ratio $\rho \in [0.05, 0.7]$, whereas admitting all queries reduces *s2t* R@1. The prediction temperature $\tau = 0.1$ achieves the highest R@1 across all evaluated settings among the displayed values; increasing it to 0.2 reduces the equal-weight average by 0.3892 points. The shared Sinkhorn temperature $\tau_s = 0.05$ yields the highest average R@1 in its sweep, although individual settings have different optima. We therefore select $\rho = 0.3$, $\tau = 0.1$, and $\tau_s = 0.05$.

Ablation Analysis. Table 5 evaluates individual components and alternative routing designs. The full model achieves an R@1 of 54.537%, exceeding the frozen model by 14.418 percentage points. Removing TSM or CRSI reduces R@1 by 2.656 and 2.499 points, respectively, supporting the use of temporal matching score calibration and structural routing for online adaptation. Removing TEM or



(a) Reliable-sample ratio ρ (b) Prediction temperature τ (c) Sinkhorn temperature τ_s

Figure 4: Sensitivity analysis of hyperparameters (ρ, τ, τ_s).

RA incurs respective drops of 1.379 and 1.376 points, indicating that confidence control and explicit representation preservation both contribute to the full framework.

The routing controls further distinguish sample exclusion from representation preservation. Under plain entropy minimization, frozen-source RA outperforms ignoring excluded queries and using a batch-entry anchor by 3.003 and 0.410 points, respectively. These comparisons favor explicitly constraining uncertain queries with a fixed source reference rather than leaving them unconstrained or anchoring them to batch-entry representations. Notably, CRSI with RA under plain entropy remains below Entropy-all, whereas the full model surpasses Entropy-all by 1.129 points. Thus, structural routing and anchoring alone do not explain the full improvement; combining them with TEM-based confidence control is important.

Removing final calibration (FC) produces the largest drop of 9.914 points, highlighting its contribution to the final retrieval rankings. Nevertheless, the full model also exceeds FC only by 6.507 points, showing that the overall improvement cannot be attributed to post-hoc calibration alone. Together, these results support combining reliability-aware online adaptation and source anchoring with final score calibration.

Table 5: Component and routing ablations.

Variant	FC	R@1	Δ
Reference settings			
Frozen backbone	No	40.119	+14.418
FC only	Yes	48.030	+6.507
Entropy-all	Yes	53.408	+1.129
Ablation variants			
w/o TSM	Yes	51.881	+2.656
w/o CRSI	Yes	52.038	+2.499
w/o TEM	Yes	53.158	+1.379
w/o RA	Yes	53.161	+1.376
w/o FC	No	44.623	+9.914
Full model	Yes	**54.537**	–
Routing controls (plain entropy)			
CRSI + ignore	Yes	50.155	+4.382
CRSI + batch anchor	Yes	52.748	+1.789
CRSI + RA	Yes	53.158	+1.379

Efficiency Analysis. As shown in Table 6, our method averages 29.65 ms/query, with a 211.8% overhead over NoTTA. This cost includes online Sinkhorn calibration, CRSI-based routing, memory-guided confidence sharpening, and frozen-source representation anchoring, while gallery embeddings remain fixed. Under the reported configurations, STAR-3D has a mean runtime close to HAT-VTR (28.39 ms/query) and lower than SAR and TCR, placing its online computational cost within the range of existing TTA methods.

Table 6: Online runtime for s2t under QS.

	Runtime (ms/query)				
Method	ScanRefer	NR3D	SceneDepict	Mean	Δ vs. NoTTA (%)
NoTTA	9.02	10.32	9.19	9.51	–
Tent	31.67	30.80	32.85	31.77	+234.1
READ	30.33	31.35	30.66	30.78	+223.7
SAR	52.51	53.80	54.95	53.75	+465.3
EATA	31.00	31.88	31.62	31.50	+231.3
TCR	72.68	74.34	73.33	73.45	+672.4
HAT-VTR	28.20	27.01	29.96	28.39	+198.6
Ours	**29.08**	**29.65**	**30.22**	**29.65**	**+211.8**

6 Conclusion

We introduced M3SP, a benchmark for evaluating the robustness of T3SR under multi-level 3D perturbations, and STAR-3D, a structure-guided test-time adaptation framework. By integrating query–gallery score calibration, reliability-aware routing, and complementary objectives for adaptation and representation preservation, STAR-3D mitigates unreliable cross-modal supervision and representation drift induced by shared query-encoder updates. Extensive experiments under query and query–gallery shifts demonstrate its effectiveness, while diagnostic analyses underscore the importance of preserving uncertain query representations alongside selective adaptation. Together, M3SP and STAR-3D provide a foundation for developing and evaluating robust cross-modal retrieval under distribution shifts.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pp. 422–440. Springer, 2020.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.

Jiaqi Chen, Daniel Barath, Iro Armeni, Marc Pollefeys, and Hermann Blum. "where am i?" scene retrieval with language. In *European Conference on Computer Vision*, pp. 201–220. Springer, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions in autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1022–1032. IEEE, 2023.

Mingtao Feng, Longlong Mei, Zijie Wu, Jianqiao Luo, Fenghao Tian, Jie Feng, Weisheng Dong, and Yaonan Wang. Partially matching submap helps: Uncertainty modeling and propagation for text to point cloud localization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8296–8305. IEEE, 2025a.

Yanglin Feng, Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, and Peng Hu. Interactive cross-modal learning for text-3d scene retrieval. *Advances in Neural Information Processing Systems*, 38:97972–97992, 2025b.

Yanglin Feng, Yang Qin, Dezhong Peng, Hongyuan Zhu, Xi Peng, and Peng Hu. Pointcloud-text matching: Benchmark dataset and baseline. *IEEE Transactions on Multimedia*, 2025c.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18941–18952. IEEE, 2022.

Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19937–19949. IEEE, 2023.

Junha Lee, Chunghyun Park, Jaesung Choe, Yu-Chiang Frank Wang, Jan Kautz, Minsu Cho, and Chris Choy. Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14089–14101. IEEE, 2025.

Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, and Mouxing Yang. Test-time adaptation for cross-modal retrieval with query shift. In *International Conference on Learning Representations*, volume 2025, pp. 76517–76544, 2025.

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. Neighborretr: Balancing hub centrality in cross-modal retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9263–9273. IEEE, 2025.

Zeyu Ma, Yuqi Li, Yizhi Luo, Xiao Luo, Jinxing Li, Chong Chen, Xian-Sheng Hua, and Guangming Lu. Discrepancy and structure-based contrast for test-time adaptive retrieval. *IEEE Transactions on Multimedia*, 26:8665–8677, 2024.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair's wmt19 news translation task submission. In *Proceedings of the fourth conference on machine translation (volume 2: shared task papers, day 1)*, pp. 314–319, 2019.

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. *arXiv preprint arXiv:2302.12400*, 2023.

Zhengxin Pan, Haishuai Wang, Fangyu Wu, Peng Zhang, and Jiajun Bu. Hubness reduction with dual bank sinkhorn normalization for cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6153–6162, 2025.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*, 2022.

Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8985–8994. IEEE, 2025.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16907–16916. IEEE, 2022.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

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. IEEE, 2022.

Wuyuqing Wang, Zeyuan Gu, and Erkun Yang. Noisy correspondence learning with modality gap direction correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10163–10171, 2026.

Xin Wei, Qin Yang, Yijie Fang, Mingrui Zhu, and Nannan Wang. 3d test-time adaptation via graph spectral driven point shift. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26762–26771. IEEE, 2025.

Qin Yang, Xin Wei, Yanjia Li, Jiasun Feng, Mingrui Zhu, Nannan Wang, and Xinbo Gao. Robust cross-modal retrieval via generative semantic refinement and exclusion-guided adaptation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yMYRr00HFS.

Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Qi Zhu, Conghui He, and Weijia Li. Where am i? cross-view geo-localization with natural language descriptions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5890–5900. IEEE, 2025.

Fuyang Yu, Zhen Wang, Dongyuan Li, Peide Zhu, Xiaohui Liang, Xiaochuan Wang, and Manabu Okumura. Towards cross-modal point cloud retrieval for indoor scenes. In *International Conference on Multimedia Modeling*, pp. 89–102. Springer, 2024.

Bingqing Zhang, Zhuo Cao, Heming Du, Yang Li, Xue Li, Jiajun Liu, and Sen Wang. Robust test-time video-text retrieval: Benchmarking and adapting for query shifts. *arXiv preprint arXiv:2604.20851*, 2026.

APPENDIX CONTENT

A More Implement Details **12**

A.1 Additional Dataset Details 12

A.2 Detailed Evaluation Protocols for Test-Time Domain Shifts 13

A.3 Details on Text Perturbations 13

A.4 Backbone details 14

A.5 Framework details 15

B M3SP Benchmark Details **17**

B.1 Micro-Level 3D Scene Perturbations 17

B.2 Meso-Level 3D Scene Perturbations 19

B.3 Macro-Level 3D Scene Perturbations 20

B.4 Further Visualization of M3SP Benchmark 21

C More Comparison Results **22**

C.1 More Hyperparameter Analysis 22

C.2 More Results on Query Shift 23

C.3 More Results on Query–Gallery Shift 24

D The Use of Large Language Models **25**

A MORE IMPLEMENT DETAILS

A.1 ADDITIONAL DATASET DETAILS

We instantiate M3SP on four ScanNet-based text–3D scene retrieval datasets: ScanRefer, Nr3D, Sr3D, and SceneDepict-3D2T. Together, these datasets cover complementary forms of language grounding, from short object-centric references to descriptions of broader scene content. We use the corresponding ScanNet scene point clouds as the 3D gallery and preserve all descriptions aligned with a scene as valid text matches. This multi-positive formulation is important for 3D-to-text retrieval, where a scene can be correctly matched to several descriptions.

ScanRefer. ScanRefer (Chen et al., 2020) is a large-scale dataset for grounding free-form language in indoor 3D scenes. It contains 51,583 human-written descriptions of 11,046 objects across 800 ScanNet scenes. Its expressions refer to target objects using attributes and spatial relations, so they provide local linguistic evidence while the retrieval target in our benchmark remains the complete scene. We use ScanRefer to evaluate robustness to natural, object-focused descriptions.

Nr3D. Nr3D (Natural Reference in 3D) (Achlioptas et al., 2020) contains 41,503 natural-language utterances collected through a two-player reference game over ScanNet scenes. The utterances identify objects, often requiring listeners to distinguish among instances of the same category by using attributes or relations. Compared with template-generated references, Nr3D supplies varied human language and offers a complementary object-centric test set for scene–text retrieval.

Sr3D. Sr3D (Spatial Reference in 3D) (Achlioptas et al., 2020) contains 83,572 synthetically generated utterances grounded in ScanNet scenes. Its template-based descriptions emphasize spatial relations between a target object and surrounding anchor objects, providing systematic coverage of

relational language. In combination with Nr3D, Sr3D lets us assess robustness across both natural and controlled object references.

SceneDepict-3D2T. SceneDepict-3D2T (Feng et al., 2025c) is a point-cloud–text matching dataset built from ScanNet scenes. In contrast to the predominantly object-level references in ScanRefer, Nr3D, and Sr3D, its descriptions characterize more of the scene and its contents. This scene-level language provides a complementary retrieval setting in which matching depends on broader geometric and semantic evidence. We include it to test whether robustness extends beyond queries centered on a single object.

A.2 Detailed Evaluation Protocols for Test-Time Domain Shifts

We evaluate robustness under two test-time domain-shift scenarios, depending on whether the shift affects only incoming queries or both retrieval modalities. In both settings, we report scene-to-text (s2t) and text-to-scene (t2s) retrieval, where the query and gallery modalities are interchanged. The target labels are used for evaluation only; adaptation operates on unlabeled target queries.

Query Shift (QS). In QS, the query distribution changes while the gallery remains clean and fixed. This models a deployed text–3D retrieval system receiving perturbed user inputs against an already indexed collection. For s2t retrieval, we corrupt the query point clouds with the 14 3D perturbation types in M3SP, each evaluated at five severity levels. For t2s retrieval, we perturb the query descriptions using character-, word-, and sentence-level transformations following TCR Li et al. (2025). Thus, each direction isolates a shift in its query modality while holding the opposite-modality gallery unchanged.

Query–Gallery Shift (QGS). QGS considers a more challenging deployment scenario in which both the query and gallery distributions differ from those of the source domain. The model must retrieve across target-domain data while adapting at test time from unlabeled queries. We study two practical forms of QGS:

- *Cross-dataset adaptation* evaluates transfer from a model fine-tuned on a labeled source dataset to a different target dataset. At test time, both queries and the corresponding gallery are drawn from the target dataset, and adaptation uses only unlabeled target queries.
- *Zero-shot adaptation* considers deployment without task-specific supervised fine-tuning. A general pretrained model is transferred directly to the target retrieval task and adapted using unlabeled target queries, with the target-domain gallery available for retrieval.

A.3 Details on Text Perturbations

We group text-query shifts into three linguistic levels—character, word, and sentence—following the taxonomy used in prior multimodal robustness work (Qiu et al., 2022). Our suite contains 15 perturbations that model distinct sources of query variation. Character-level operations simulate recognition or typing errors: optical character recognition (OCR) noise applies visually plausible character confusions, while character insertion, replacement, swapping, and deletion simulate extra, incorrect, transposed, or omitted characters. Word-level operations modify lexical choice or ordering. Synonym replacement selects non-stopwords and substitutes WordNet synonyms; random insertion adds a synonym of a sampled source word; random swap exchanges word positions; and random deletion removes words. These EDA-based operations first lowercase the query and normalize punctuation and other nonalphabetic characters to spaces. Punctuation insertion is applied separately by adding punctuation marks before selected words. Sentence-level operations preserve the intended scene reference while changing its expression through formal or casual style, passive or active voice, or back-translation. Together, the three levels test whether retrieval remains stable when object and attribute terms are corrupted, lexical content or order changes, or the same intent is phrased differently.

We use severity 7 for character-level perturbations, severity 2 for word-level perturbations, and severity 1 for sentence-level perturbations. At these settings, the character-operation rates are 0.15 for OCR and 0.10 for each other character operation; the word-operation rates are 0.10, except for punctuation insertion at 0.20. These are calibrated rates for our short, information-dense referring queries. Their meanings depend on the operator: character rates specify the proportion of words

Table 7: A representative source query and its generated text-perturbation variants. Orange text highlights perturbed spans; for deletions, it marks the affected token. Examples are shown at the reported severity levels: character 7, word 2, and sentence 1.

Perturbation Method	**Abbr.**	**Text query**
Clean Source		
Original	–	this is a round table. it is in between two couches.
Character-Level		
Optical Character Recognition	OCR	this is a round table. it is in between tw0 couches.
Character Insertion	CI	thsiKs is a round Wtablve. it is in between two couches.
Character Replacement	CR	t($s is a round I9ble. it is in between two couches.
Character Swap	CS	htsi is a round table. it is in tbewene two couches.
Character Deletion	CD	th is a round table. it is in btwn two couches.
Word-Level		
Synonym Replacement	SR	this is a round table it is in between two couch
Random Insertion	RI	this is a round table it is equal in between two couches
Random Swap	RS	it is a round table this is in between two couches
Random Deletion	RD	this is a round table it between two couches
Punctuation Insertion	IP	this is a round table. it is in between , two , couches.
Sentence-Level		
Formal Style	Formal	This is a round table, and it is in between two couches.
Casual Style	Casual	round table in between 2 couches
Passive Voice	Passive	a round table is used to be used by this and is in between two couches.
Active Voice	Active	the round table is in between two couches.
Back-Translation	Backtrans	This is a round table. It's between two sofas.

selected for character edits; synonym replacement, random insertion, and random swap use the rate to determine the number of edits; random deletion uses it as the per-word deletion probability; and punctuation insertion uses it to determine how many word positions receive punctuation. Thus, equal rate values across operations do not imply equal perceptual severity.

For word-level edits and sentence-level rewrites, we use MPNet cosine similarity to filter candidates and retain only outputs scoring at least 0.90 against the source query. We try at most 100 candidates and fall back to the original query if none passes; character-level edits are not subject to this semantic filter because they are intended to introduce orthographic noise. We generate the four style and voice variants with Styleformer. For back-translation, we translate English queries into German and back using Facebook FAIR's WMT19 English–German and German–English models (Ng et al., 2019). We first try to translate the full query, then use sentence-wise translation if the full-query candidate fails validation. The final candidate must pass the semantic-similarity threshold and structural checks; otherwise, we retain the original query. Table 7 shows one clean source query and a seed-fixed output for each perturbation.

A.4 Backbone details

CrossOver. We initialize CrossOver from its `scene_crossover_scannet+scan3r.pth`[1] checkpoint and initialize the text encoder with BLIP-large `model_large.pth` weights. Its 3D tower uses a sparse ResNet-34 to encode voxelized scene coordinates and RGB features, followed by an MLP that maps 512-dimensional scene features to the 768-dimensional retrieval space. The text tower uses a BLIP MED/BERT encoder and projects its [CLS] representation to the same space with a separate MLP. Across these four datasets, we fine-tune the two towers and their projection heads on clean scene–description pairs from that dataset's training split. A symmetric multi-positive contrastive loss treats every description of a scene as a valid match, and the checkpoint is selected using the corresponding development split. The resulting dataset-specific model supplies the frozen source parameters for test-time adaptation.

[1] `https://github.com/GradientSpaces/CrossOver`

Mosaic3D. We use the `sc+ar+sc+++ma+st.ckpt`[2] Mosaic3D checkpoint with its corresponding Recap-CLIP text-encoder weights. The 3D tower uses a ScanNet-conditioned sparse U-Net (SpUNet). We mean-pool its final 768-dimensional voxel features to form a scene embedding. The text tower is a 24-layer Recap-CLIP Transformer with a 768-dimensional output, placing both modalities in a shared retrieval space. For each of the same four datasets, we keep both pretrained towers fixed and fit a dataset-specific scene-side alignment head on clean training pairs. This head subtracts the training-scene feature mean before an identity-initialized, bias-free 768×768 projection; its output is ℓ_2-normalized and trained with a bidirectional multi-positive contrastive loss. We select the head using development-set retrieval and refit it on the full training split before test-time adaptation.

A.5 Framework details

The stream-level procedure of STAR-3D is summarized in Algorithm 1. The details below further specify the calibration, routing, and optimization steps.

Online calibration. The gallery embeddings are indexed once, and the online query encoder is initialized from the source model; only the configured query-side parameters are optimized. The optimizer and both memories persist across target batches, whereas the saved source parameters provide a fixed reference throughout adaptation. Temporal Similarity Memory (TSM) retains at most $K-1$ detached batches of post-adaptation *raw* scores, including partial batches with different row counts. Given nonempty history, online calibration concatenates these scores with the current raw score matrix and applies Eq. (4) in the log domain. For numerical stability, we initialize

$$Z_{ij}^{(0)} = \frac{\bar{\mathcal{S}}_{t,ij} - \max_k \bar{\mathcal{S}}_{t,ik}}{\tau_s},$$

and alternately subtract row-wise and column-wise $\mathrm{logsumexp}$ terms. Exponentiating the resulting matrix yields the scaled scores, from which only the rows corresponding to the current batch are retained. When TSM is empty, the raw scores are directly used for candidate selection. Since a batch enters TSM only after its adaptation steps are completed, neither candidate selection nor optimization has access to future queries. For rectangular score matrices, the scaled values are used as relative matching scores rather than as a strict one-to-one assignment.

CRSI routing and entropy memory. Each row of $\hat{\mathcal{S}}_t$ determines one provisional gallery candidate, and the same gallery item may be selected by multiple queries. Centered Residual Structural Inconsistency (CRSI) is computed using the mean of these *selected* gallery features, including repeated candidates, rather than the mean of the full gallery. For fixed candidate assignments, let

$$\mathbf{h}_{t,i}^{Q} = \boldsymbol{\mu}_Q + \mathbf{u}_i, \qquad \mathbf{c}_{t,i} = \boldsymbol{\mu}_G + \mathbf{v}_i.$$

After centering, their residual difference becomes

$$(\mathbf{u}_i - \bar{\mathbf{u}}) - (\mathbf{v}_i - \bar{\mathbf{v}}),$$

so a common translation in either modality is canceled in Eq. (7). The resulting CRSI partition $(\mathcal{R}_t, \mathcal{U}_t)$ is detached from the online encoder and remains fixed throughout adaptation of the current batch.

Before the first optimization step, Temporal Entropy Memory (TEM) receives detached pre-adaptation entries from $\mathcal{R}_t$. Each entry stores the selected query–gallery pair features, its CRSI value, and its candidate-set entropy, although only the entropy is used to determine the threshold in Eq. (10). TEM follows a FIFO policy and discards the oldest entries once its capacity is reached. Since CRSI routing always retains at least one reliable query, TEM is nonempty when its maximum retained entropy defines the batch-wise threshold $\bar{E}_t$. If all newly admitted entries remain in memory and the pre-adaptation encoder is deterministic, every query in $\mathcal{R}_t$ initially satisfies the hard entropy gate, while their contributions are differentiated by $W_{t,i}$. As the online encoder is updated, entropy is recomputed and a query may subsequently move from entropy sharpening to source anchoring.

[2] `https://github.com/NVlabs/Mosaic3D`

Algorithm 1 The overall framework workflow of STAR-3D

Input: Source model f_{Θ_s}; fixed gallery embeddings $\mathbf{H}^G$; Ordered query batches $\{\mathcal{X}_t^Q\}_{t=1}^T$; adaptation steps N_{tta}; Calibration and optimization hyperparameters
Output: Retrieval score matrix $\mathcal{S}^{\text{out}}$

Initialize $\Theta \leftarrow \Theta_s$ and AdamW for query-side parameters θ
Initialize TSM $\mathcal{M}_0^S \leftarrow \varnothing$ and TEM $\mathcal{M}_0^E \leftarrow \varnothing$
Initialize output matrix $\mathcal{S}^{\text{out}}$
for $t = 1, \ldots, T$ **do**
  Initial scoring and online calibration
  $\mathbf{H}_t^Q \leftarrow f_\Theta^Q(\mathcal{X}_t^Q)$
  $\mathbf{H}_t^S \leftarrow \text{sg}\Big(f_{\Theta_s}^Q(\mathcal{X}_t^Q)\Big)$
  $\mathcal{S}_t \leftarrow \mathbf{H}_t^Q(\mathbf{H}^G)^\top$
  if $\mathcal{M}_{t-1}^S \neq \varnothing$ **then**
    $\hat{\mathcal{S}}_t \leftarrow \text{TSMCalibrate}(\mathcal{S}_t, \mathcal{M}_{t-1}^S)$ ▷ Eq. (4)
  else
    $\hat{\mathcal{S}}_t \leftarrow \mathcal{S}_t$
  end if
  Provisional correspondences and CRSI routing
  $j_{t,i}^\star \leftarrow \arg\max_j \hat{\mathcal{S}}_{t,ij}$, $\mathbf{c}_{t,i} \leftarrow \mathbf{h}_{j_{t,i}^\star}^G$ for each i
  $\mathbf{C}_t \leftarrow [\mathbf{c}_{t,1}, \ldots, \mathbf{c}_{t,B}]^\top$
  $(\mathcal{R}_t, \mathcal{U}_t) \leftarrow \text{CRSIRoute}(\mathbf{H}_t^Q, \mathbf{C}_t)$ ▷ Eq. (8)
  Pre-adaptation TEM update
  $\widehat{\mathbf{p}}_t^{(0)} \leftarrow \text{Softmax}\Big(\mathbf{H}_t^Q\mathbf{C}_t^\top/\tau\Big)$
  $E_{t,i}^{(0)} \leftarrow -\sum_{j=1}^B \widehat{p}_{t,ij}^{(0)} \log \widehat{p}_{t,ij}^{(0)}$ for each i
  $\mathcal{M}_t^E \leftarrow \text{FIFO}\Big(\mathcal{M}_{t-1}^E, \{\text{sg}(E_{t,i}^{(0)})\}_{i\in\mathcal{R}_t}\Big)$
  $\bar{E}_t \leftarrow \max_{E\in\mathcal{M}_t^E} E$ ▷ Eq. (10)
  Online adaptation
  for $k = 1, \ldots, N_{\text{tta}}$ **do**
    $\mathbf{H}_t^{Q,(k)} \leftarrow f_\Theta^Q(\mathcal{X}_t^Q)$
    $\widehat{\mathbf{p}}_t^{(k)} \leftarrow \text{Softmax}\Big(\mathbf{H}_t^{Q,(k)}\mathbf{C}_t^\top/\tau\Big)$
    $E_{t,i}^{(k)} \leftarrow -\sum_{j=1}^B \widehat{p}_{t,ij}^{(k)} \log \widehat{p}_{t,ij}^{(k)}$ for each i
    $\mathcal{A}_t^{(k)} \leftarrow \{i \in \mathcal{R}_t : E_{t,i}^{(k)} \leq \bar{E}_t\}$
    $\mathcal{P}_t^{(k)} \leftarrow \{1, \ldots, B\} \setminus \mathcal{A}_t^{(k)}$
    $W_{t,i}^{(k)} \leftarrow \left[1 - \frac{\text{sg}(E_{t,i}^{(k)})}{\bar{E}_t + \epsilon}\right]_+$ for each i
    Compute $\mathcal{L}_{\text{MACS}}^{(k)}$ on $\mathcal{A}_t^{(k)}$ using $W_{t,i}^{(k)}$ ▷ Eq. (11)
    Compute $\mathcal{L}_{\text{RA}}^{(k)}$ on $\mathcal{P}_t^{(k)}$ using $\mathbf{H}_t^S$ ▷ Eq. (13)
    $\mathcal{L}^{(k)} \leftarrow \lambda_E\mathcal{L}_{\text{MACS}}^{(k)} + \lambda_A\mathcal{L}_{\text{RA}}^{(k)}$ ▷ Eq. (14)
    Update only θ with one AdamW step on $\mathcal{L}^{(k)}$
  end for
  Post-adaptation scoring and TSM update
  $\mathbf{H}_t^{Q,+} \leftarrow f_\Theta^Q(\mathcal{X}_t^Q)$
  $\mathcal{S}_t^+ \leftarrow \mathbf{H}_t^{Q,+}(\mathbf{H}^G)^\top$
  Write $\mathcal{S}_t^+$ to the corresponding rows of $\mathcal{S}^{\text{out}}$
  $\mathcal{M}_t^S \leftarrow \text{FIFO}\big(\mathcal{M}_{t-1}^S, \text{sg}(\mathcal{S}_t^+)\big)$
end for
if final calibration is enabled **then**
  $\mathcal{S}^{\text{out}} \leftarrow \text{FC}(\mathcal{S}^{\text{out}})$
end if
return $\mathcal{S}^{\text{out}}$

Routed optimization and final calibration. At each optimization step k, the online encoder recomputes the query representations and candidate-set entropies while the provisional gallery candidates, the CRSI partition $(\mathcal{R}_t, \mathcal{U}_t)$, and the threshold $\bar{E}_t$ remain fixed. The dynamically determined admission set

$$\mathcal{A}_t^{(k)} = \left\{ i \in \mathcal{R}_t \mid E_{t,i}^{(k)} \leq \bar{E}_t \right\}$$

receives weighted entropy minimization, whereas its complement

$$\mathcal{P}_t^{(k)} = \{1, \ldots, B_t\} \setminus \mathcal{A}_t^{(k)}$$

receives frozen-source representation anchoring. Consequently, every query routed to $\mathcal{U}_t$ is always assigned to the anchoring branch, while a CRSI-reliable query can switch from sharpening to anchoring if its entropy exceeds the fixed threshold during adaptation.

The entropy weight is computed from detached entropy as

$$W_{t,i}^{(k)} = \max\left(1 - \frac{\mathrm{stopgrad}(E_{t,i}^{(k)})}{\bar{E}_t + \epsilon}, 0\right), \qquad \epsilon = 10^{-3}.$$

Notably, when $E_{t,i}^{(k)} = \bar{E}_t$, the weight equals $\epsilon/(\bar{E}_t + \epsilon)$ rather than zero. The source representation is obtained by encoding the same target query with the saved source parameters and is detached from optimization; therefore, the anchoring term introduces no gallery pseudo-label. If either branch is empty, its corresponding loss contributes zero. AdamW updates only the configured query-side parameters, and its optimizer state is carried across target batches.

After completing the adaptation steps for batch t, the online encoder produces the post-adaptation raw score matrix $\mathcal{S}_t^+$. These scores are written to $\mathcal{S}^{\mathrm{out}}$ and a detached copy is appended to TSM for subsequent batches. When enabled, final calibration (FC) applies a separate Sinkhorn calibration pass to the completed output matrix before retrieval ranking, potentially using calibration settings different from those of online TSM. FC is used only for final evaluation and does not feed back into either memory or any online parameter update.

B M3SP Benchmark Details

M3SP (**M**ulti-level **3**D **S**cene **P**erturbations) evaluates the robustness of T3SR models to controlled perturbations of 3D queries while keeping the text gallery and scene-level retrieval targets unchanged. It comprises 14 perturbation types, each instantiated at five progressively stronger severity levels, yielding 70 conditions in total. We organize these perturbations according to the spatial extent of the evidence they affect: *Micro-Level* perturbations alter individual point measurements, *Meso-Level* perturbations act on coherent object instances, and *Macro-Level* perturbations disrupt scene-wide acquisition patterns, spatial registration, or geometry–appearance correspondence.

Each input is represented as an axis-aligned colored point cloud with instance assignments,

$$P = \{(\mathbf{x}_i, \mathbf{c}_i, \ell_i)\}_{i=1}^{N},$$

where $\mathbf{x}_i \in \mathbb{R}^3$ denotes the 3D coordinate in meters, $\mathbf{c}_i \in [0,1]^3$ denotes the RGB value, and ℓ_i denotes the instance assignment. Unassigned background points are excluded before corruption. Instance assignments remain attached to surviving points and are used only by perturbations requiring object-level sampling; the assignments themselves are never modified. Unless otherwise specified, geometric perturbations preserve RGB, appearance perturbations preserve XYZ and point count, whereas removal and injection operators change the number of points. Table 8 summarizes the 14 perturbations and their abbreviations, while Table 9 reports the complete parameter settings for all five severity levels.

B.1 Micro-Level 3D Scene Perturbations

Micro-Level perturbations model errors in individual point measurements and isolated false returns. They operate directly on the scene point set without relying on instance identity and therefore do not explicitly target any semantic object. The Gaussian, Uniform, and Impulse formulations follow

Table 8: Overview of the M3SP perturbations. The 14 perturbations are grouped by the spatial extent of the corrupted evidence; abbreviations are used throughout the experimental tables.

Perturbation	Abbr.	Description
		Micro-Level
Gaussian Noise	GN	Adds dense, zero-mean coordinate jitter to model continuous ranging uncertainty.
Uniform Noise	UN	Adds bounded, independent coordinate offsets to model measurement uncertainty with a finite range.
Impulse Noise	IN	Applies sparse, large signed coordinate jumps to selected points.
Outlier Injection	OI	Adds isolated false returns sampled around the scene bounds.
		Meso-Level
Incomplete Echo	IE	Removes spatially dispersed returns within individual object instances.
Object Cutout	OC	Removes a contiguous local surface region from each eligible object instance.
Local Deformation	LD	Smoothly warps a local object region while preserving point count.
Object Color Shift	OCS	Coherently alters the hue and saturation of individual object instances.
		Macro-Level
Beam Missing	BM	Removes structured horizontal bands across the scene to model scan-band dropout.
FOV Lost	FL	Removes a horizontal angular sector to simulate restricted scene coverage.
Registration Drift	RD	Introduces smooth spatially varying translation and yaw drift across the scene.
Texture Smoothing	TSm	Suppresses local appearance details through neighborhood-based color smoothing.
Texture Shuffle	TSh	Disrupts the local spatial organization of colors while largely preserving their distribution.
RGB–XYZ Misalignment	RXM	Breaks geometry–appearance correspondence by assigning colors from spatially displaced locations.

established 3D sensor-corruption designs (Kong et al., 2023; Dong et al., 2023; Kar et al., 2022), while their severity ranges are calibrated for our retrieval setting.

Gaussian Noise. Gaussian Noise models dense random ranging uncertainty by independently perturbing the three coordinates of every point:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_3), \tag{15}$$

where σ controls the magnitude of the coordinate jitter.

Uniform Noise. Uniform Noise models bounded measurement uncertainty. For each point and coordinate dimension, the perturbation is independently sampled as

$$\tilde{x}_{ij} = x_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{U}(-a, a), \tag{16}$$

where a specifies the maximum absolute offset along each coordinate dimension. Unlike Gaussian Noise, a denotes a hard perturbation bound rather than a standard deviation.

Impulse Noise. Impulse Noise models sparse and abrupt measurement spikes. A severity-dependent subset of points is sampled without replacement, and each selected coordinate receives an independently signed fixed-magnitude jump:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + 0.08\,\mathbf{s}_i, \qquad \mathbf{s}_i \in \{-1, +1\}^3. \tag{17}$$

The perturbation magnitude remains fixed at 8 cm per axis, while severity controls the fraction of affected points.

Outlier Injection. Outlier Injection models isolated false or floating returns caused by erroneous depth measurements or mismatches. Given the scene bounding box, we sample a severity-dependent number of additional points from a box expanded by 0.75% along each axis extent. Each injected point is assigned the scene-mean RGB value to avoid introducing an unrelated appearance anomaly. The RGB values of all original points remain unchanged.

Table 9: Perturbation parameter specifications across severity levels. Values correspond to the effective settings used in M3SP; perturbations with multiple parameters span multiple rows for readability.

Perturbation	Parameter	S1	S2	S3	S4	S5
	Micro-Level					
Gaussian Noise	Standard deviation σ (cm)	0.45	0.75	1.05	1.35	1.80
Uniform Noise	Half-width a (cm)	0.5	1	2	3.5	5
Impulse Noise	Selected points (%)	2.67	3.20	4	5.33	8
	Jump magnitude per axis (cm)	8	8	8	8	8
Outlier Injection	Added points (%)	0.0375	0.05625	0.075	0.1125	0.15
	Bounding-box padding (%)	0.75	0.75	0.75	0.75	0.75
	Meso-Level					
Incomplete Echo	Removed points per instance (%)	5	7.5	10	12.5	15
Object Cutout	Local removal per instance (%)	2	3.5	5	6.5	8
Local Deformation	Selected local points per instance (%)	15	20	25	30	35
	Displacement ratio ρ of object diagonal (%)	1	1.5	2	2.5	3
	Absolute displacement cap (cm)	4	4	4	4	4
Object Color Shift	Hue-shift magnitude (°)	10	17.5	25	32.5	40
	Saturation multiplier	1.05	1.10	1.15	1.20	1.25
	Macro-Level					
Beam Missing	Occupied bins removed (%)	10	15	20	25	30
	Number of equal-height bins	64	64	64	64	64
FOV Lost	Retained horizontal FOV (°)	315	306	297	288	279
	Removed horizontal wedge (°)	45	54	63	72	81
Registration Drift	Number of principal-axis segments	4	4	5	5	6
	Endpoint translation cap (cm)	1.5	2.25	3	4	5
	Endpoint yaw cap (°)	0.4	0.6	0.8	1.0	1.2
	Per-point displacement cap (cm)	3	4.5	6	8	10
Texture Smoothing	Number of neighbors k	8	12	16	20	24
	Color blending weight α	0.30	0.40	0.50	0.575	0.65
Texture Shuffle	Selected points per instance (%)	25	37.5	50	62.5	75
RGB–XYZ Misalignment	Selected points per instance (%)	30	45	60	72.5	85
	Lookup offset (cm)	2	3	4	5	6

B.2 Meso-Level 3D Scene Perturbations

Meso-Level perturbations exploit instance assignments to model degradations localized to coherent object surfaces or appearance. Removal and deformation operators are applied only to eligible instances containing at least 10 points, whereas Object Color Shift is applied to every assigned foreground instance.

Incomplete Echo. Incomplete Echo models incomplete object returns caused by occlusion, weak reflectance, or viewpoint changes. For each eligible instance, a severity-dependent fraction of points is sampled uniformly without replacement and removed. The resulting missing points are spatially dispersed over the object surface, degrading instance completeness while largely preserving object identity.

Object Cutout. Object Cutout instead models a localized surface failure. A point is uniformly sampled from each eligible instance as the cutout center, after which the prescribed number of nearest points within that instance is removed. This produces a single contiguous missing region and is therefore complementary to the dispersed missing pattern introduced by Incomplete Echo.

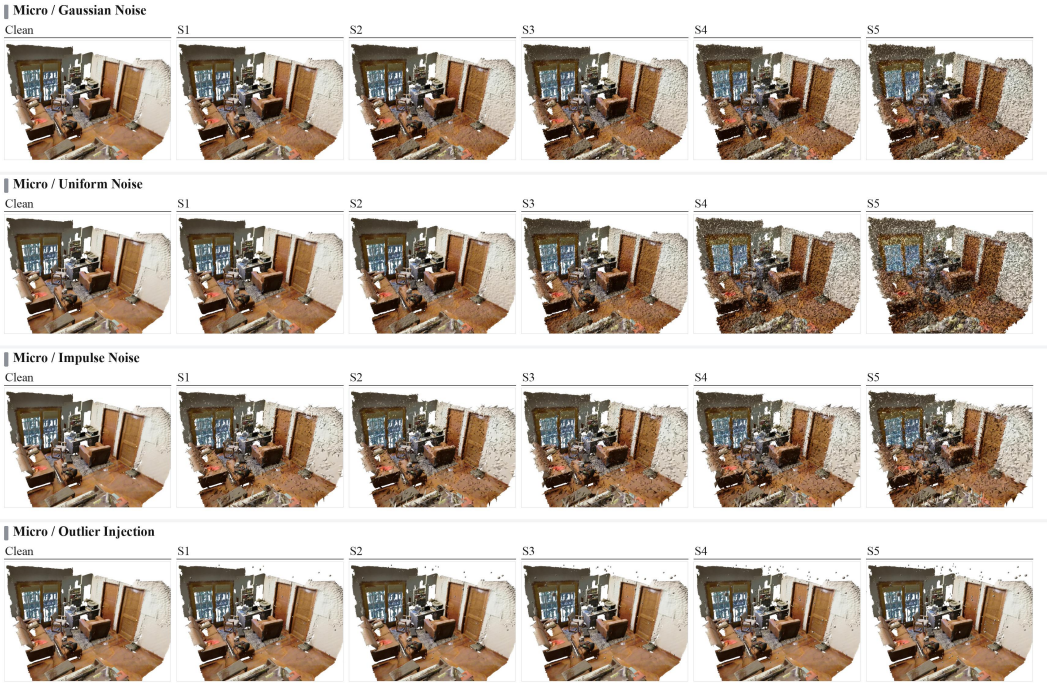


Figure 5: Visualization Examples of Different Micro-level 3D Scene Perturbations

Local Deformation. Local Deformation models local reconstruction errors or mild non-rigid surface distortions. For each eligible instance, a center point is sampled and a severity-dependent fraction of its nearest points is selected. Let d_i denote the distance of a selected point from the center, R the largest distance within the selected region, and $\mathbf{u}$ a random unit direction. The displacement is defined as

$$\Delta \mathbf{x}_i = m \exp\left[-2(d_i/R)^2\right] \mathbf{u}, \qquad m = \min(\rho d_{\text{obj}}, 4\ \text{cm}), \tag{18}$$

where d_{obj} denotes the instance bounding-box diagonal and ρ controls the severity-dependent displacement magnitude. The deformation therefore decreases smoothly from the sampled center toward the region boundary while preserving point count, RGB values, and instance assignments.

Object Color Shift. Object Color Shift targets instance-level appearance rather than geometry. For each foreground instance, RGB values are converted to HSV space, a coherently signed hue shift is applied, and saturation is scaled by a severity-dependent factor before conversion back to RGB. The operation changes object appearance while leaving XYZ coordinates and point count unchanged.

B.3 Macro-Level 3D Scene Perturbations

Macro-Level perturbations affect scene-wide acquisition patterns, spatial registration, or geometry–appearance correspondence. Unlike Micro-Level perturbations, they introduce structured changes whose effects depend on the global scene layout or on spatial relationships among points.

Beam Missing. Beam Missing models structured scan-band dropout. The vertical extent of the scene is divided into 64 equal-height bins, and a severity-dependent fraction of occupied bins is sampled and removed in its entirety. This produces coherent band-shaped missing regions across the scene rather than independent point dropout.

FOV Lost. FOV Lost follows the sector-cropping form used in established 3D corruption benchmarks. Each point is expressed by its horizontal angle around the scene's XY centroid, and only points within a retained angular sector centered on the positive x direction are preserved. Increasing severity progressively narrows the retained field of view, modeling restricted scene coverage caused by limited sensing range or sensor obstruction. The exact retained FOV values are calibrated for our retrieval setting.



Figure 6: Visualization Examples of Different Meso-level 3D Scene Perturbations

Registration Drift. Registration Drift models accumulated pose errors during multi-frame fusion or global scene registration. We first estimate the dominant principal axis of the scene and partition it into a small number of ordered segments. Translation and yaw offsets are defined at segment anchors and smoothly interpolated along the principal axis, yielding a spatially varying transformation across the scene. The final displacement of each point is bounded by a severity-dependent cap, while RGB values and point count remain unchanged.

Texture Smoothing. Texture Smoothing models the loss of local appearance detail while preserving geometry. Within each object instance, we identify up to k nearest neighbors in XYZ space and compute an inverse-distance-weighted local color estimate $\bar{\mathbf{c}}_i$. The point color is then updated as

$$\tilde{\mathbf{c}}_i = (1 - \alpha)\mathbf{c}_i + \alpha\bar{\mathbf{c}}_i, \tag{19}$$

where k and α jointly control the smoothing strength. XYZ coordinates and point count remain unchanged.

Texture Shuffle. Texture Shuffle preserves geometry and the overall set of colors within each instance while disrupting their local spatial organization. A severity-dependent subset of instance points is sampled without replacement, and their RGB values are randomly permuted among the selected locations. This retains the selected color distribution but weakens the correspondence between local appearance and spatial position.

RGB–XYZ Misalignment. RGB–XYZ Misalignment directly perturbs geometry–appearance correspondence. Within each instance, a severity-dependent subset of points is selected. For every selected point, a lookup location is displaced along a random direction, and the RGB value of a nearby same-instance point at that displaced location is assigned to the original point. Consequently, geometry and point count remain unchanged, while appearance is deliberately transferred from a nearby but incorrect spatial location.

B.4 Further Visualization of M3SP Benchmark

Figures 5, 6, and 7 use a representative scene to illustrate how M3SP perturbations appear at the micro, meso, and macro levels. Within each illustrated operator, the clean example is followed by its five severity settings. The micro-level panels show changes to individual point measurements,

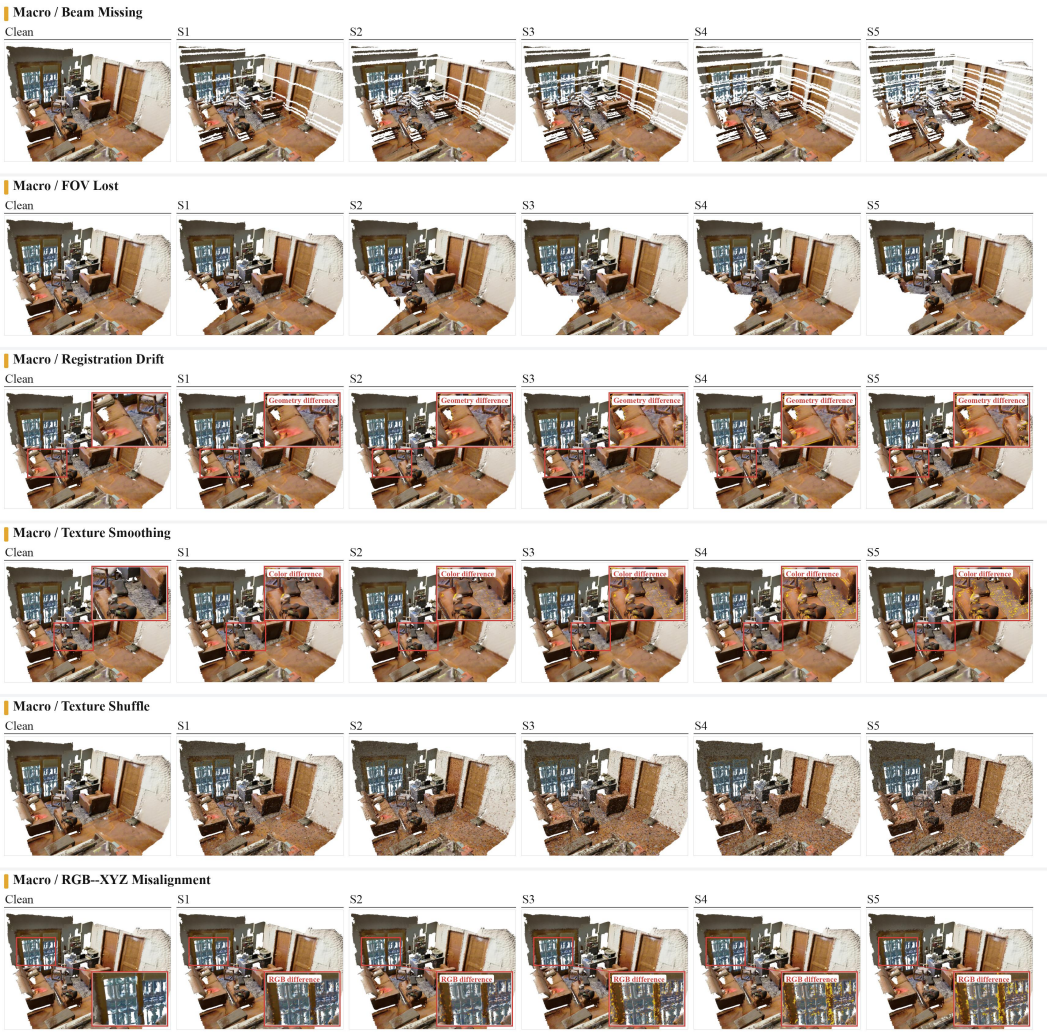


Figure 7: Visualization Examples of Different Macro-level 3D Scene Perturbations

the meso-level panels highlight changes concentrated on object regions, and the macro-level panels show broader changes to scene coverage, registration, or geometry–appearance correspondence. The side-by-side examples make the spatial character and within-operator progression of these perturbations easier to inspect alongside the parameter specifications in Table 9.

C More Comparison Results

C.1 More Hyperparameter Analysis

We study five inference-time hyperparameters with query-side normalization adaptation on ScanRefer and SceneDepict in both retrieval directions (Figs. 8 and 9). Performance is relatively insensitive to the two memory sizes: sweeping entropy-memory capacity from 8 to 128 and online Sinkhorn history from 1 to 128 batches changes the equally weighted mean R@1 across the four settings by 0.08 and 0.13 percentage points, respectively. Histories of 64 or 128 batches exceed the approximately 22–23 batches in an s2t stream and provide no consistent gain. Increasing the shared online and final Sinkhorn iteration count from 1 to 5 improves mean R@1 by 1.02 points; increasing it from 5 to 20 adds only 0.18 points. Accordingly, we use an entropy-memory capacity of 128, a 32-batch online history (between the evaluated values of 16 and 64), and five iterations in each Sinkhorn stage. Learning-rate sensitivity varies more across settings. In three paired runs, rates of 3×10^{-5}

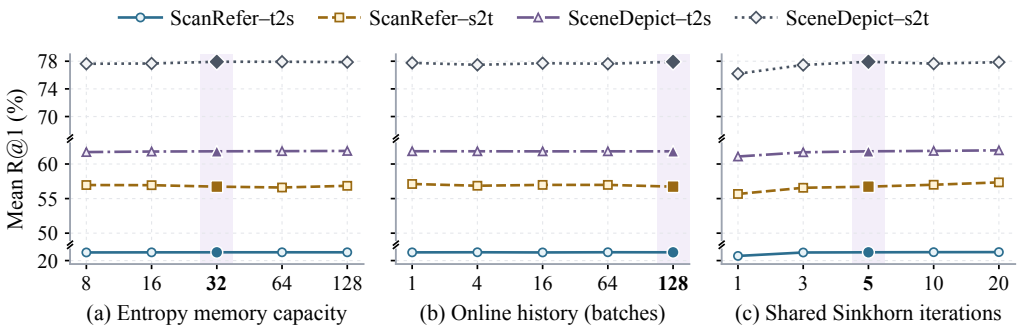


Figure 8: Sensitivity to entropy-memory capacity, online Sinkhorn history length, and the shared online/final Sinkhorn iteration count. Curves show R@1 averaged over perturbations for ScanRefer and SceneDepict in both retrieval directions; shading marks the reference value in each sweep.

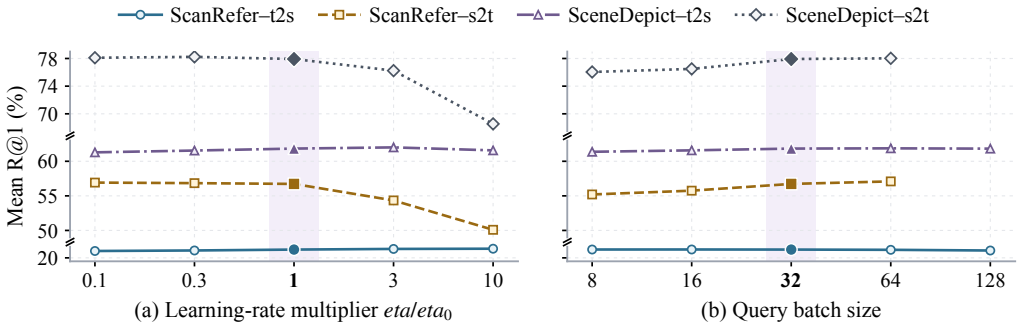


Figure 9: Sensitivity to the learning-rate multiplier and query batch size. Curves use the same datasets, retrieval directions, and R@1 averaging as Fig. 8; shading marks the reference values. The s2t points at batch size 128 are omitted after an out-of-memory run.

rather than 3×10^{-6} for ScanRefer t2s and 9×10^{-4} rather than 3×10^{-3} for SceneDepict s2t improve R@1 by 0.177 and 0.564 points, respectively, with positive differences in every run. The configured rates for ScanRefer t2s/s2t and SceneDepict t2s/s2t are 3×10^{-5}, 1×10^{-3}, 3×10^{-5}, and 9×10^{-4}. For batch size, 64 is the best valid value in both s2t settings, improving over 32 by 0.640 and 0.517 points; the selected t2s alternatives improve by only 0.013 and 0.075 points. We retain a common batch size of 32 because batch size also changes update frequency and history coverage. The query-batch-size-128 points are omitted from both s2t curves after an out-of-memory run. The paired-run comparisons are descriptive, since candidate values were selected using one of those runs.

C.2 More Results on Query Shift

We further evaluate query shift on SceneDepict, Nr3D, and Sr3D with both CrossOver and Mosaic3D. The six tables report text-to-scene (t2s) retrieval under 15 text perturbations and scene-to-text (s2t) retrieval under 14 3D scene perturbations, with R@1 averaged uniformly over the conditions in each task.

For t2s retrieval, Tables 16, 17, and 18 show strong but not uniform gains. On SceneDepict, STAR-3D reaches average R@1 scores of 62.17% with CrossOver and 22.88% with Mosaic3D, compared with 60.82% for EATA and 20.49% for TCR, the respective strongest alternatives. On Nr3D, the corresponding scores are 24.59% versus 22.85% for TCR and 9.17% versus 8.25% for HAT-VTR. Our method leads on every listed text perturbation for both backbones on these two datasets. The Sr3D result is more mixed: STAR-3D improves the Mosaic3D average to 2.50% from 2.23% without adaptation, but its CrossOver average of 10.39% is slightly below TCR's 10.49%.

For s2t retrieval, Tables 19, 20, and 21 show the highest average R@1 for STAR-3D on all three datasets with both backbones. Against HAT-VTR, the strongest alternative in each case, STAR-3D reaches 76.73% versus 72.14% with CrossOver and 37.22% versus 31.63% with Mosaic3D on SceneDepict; 51.20% versus 45.41% and 15.65% versus 13.56% on Nr3D; and 11.20% versus 9.01% and 3.69% versus 2.67% on Sr3D. The average gains extend across the micro-, meso-, and macro-level corruption groups, although individual Sr3D conditions still favor a baseline. Since the fixed Sr3D corpus includes original training scenes, its results measure corruption robustness rather than scene-disjoint generalization.

C.3 More Results on Query–Gallery Shift

C.3.1 More Results on Zero-Shot Adaptation

Zero-shot QGS starts from the original pretrained Mosaic3D model without target-dataset fine-tuning. Table 10 shows that STAR-3D attains or ties the highest R@1 in all four retrieval settings. On Sr3D, STAR-3D and READ both reach 1.71% on s2t, while STAR-3D obtains 1.66% on t2s versus 1.24% for READ. On Nr3D, STAR-3D reaches 3.27% on s2t and 1.75% on t2s, compared with 2.31% and 1.42% for READ. These fixed-corpus results show gains when starting from pretrained weights; the Sr3D corpus includes original training scenes, so they do not establish scene-disjoint generalization.

Table 10: Results of *Zero-shot Adaptation* on Sr3D and Nr3D.

	Sr3D		**Nr3D**	
Method	*s2t* R@1↑	*t2s* R@1↑	*s2t* R@1↑	*t2s* R@1↑
Mosaic3D	1.14	0.92	1.85	1.38
+ Tent	1.42	0.43	1.78	0.42
+ READ	**1.71**	1.24	2.31	1.42
+ SAR	1.14	0.92	1.97	1.21
+ EATA	1.14	0.92	1.97	1.21
+ TCR	0.43	0.48	0.68	0.77
+ HAT-VTR	0.28	0.46	0.71	0.89
+ Ours	**1.71**	**1.66**	**3.27**	**1.75**

C.3.2 More Results on Cross-Dataset Adaptation

We evaluate cross-dataset QGS on five additional dataset pairs in both transfer directions and both retrieval tasks. For ScanRefer and Nr3D (Table 11), our method leads all four Mosaic3D comparisons, including 7.02 versus 6.27 for HAT-VTR on ScanRefer → Nr3D s2t. With CrossOver, it leads on Nr3D → ScanRefer, reaching 13.56 and 13.16 on s2t and t2s, respectively, but HAT-VTR and TCR lead on the corresponding ScanRefer → Nr3D tasks.

The ScanRefer–Sr3D results (Table 12) are similarly asymmetric. On Sr3D → ScanRefer, our method leads in both tasks with both backbones; for example, it reaches 7.09 versus 5.92 for HAT-VTR on Mosaic3D s2t and 12.18 versus 11.67 for EATA on CrossOver t2s. In the opposite direction, it leads only on Mosaic3D s2t, while the strongest baselines retain the advantage in the other three comparisons.

For Nr3D and SceneDepict (Table 13), the gains are most pronounced when SceneDepict is the target. On Nr3D → SceneDepict s2t, our method reaches 23.24 versus 18.83 for HAT-VTR with CrossOver and 22.33 versus 17.73 with Mosaic3D; it also leads in both t2s comparisons. The reverse transfer yields smaller gains, and EATA exceeds our method on CrossOver t2s (18.26 versus 17.92).

The Nr3D–Sr3D pair (Table 14) shows more modest and mixed improvements. On Sr3D → Nr3D, our method reaches 12.75 versus 11.28 for HAT-VTR on CrossOver s2t and 13.81 versus 13.62 for EATA on CrossOver t2s, whereas HAT-VTR remains ahead on Mosaic3D s2t (4.24 versus 4.13). On Nr3D → Sr3D t2s, SAR leads our method with CrossOver (11.34 versus 10.83).

SceneDepict and Sr3D exhibit the largest directional contrast (Table 15). On Sr3D → SceneDepict, our method leads with both backbones and in both tasks, including 24.53 versus 17.71 for HAT-VTR on CrossOver t2s and 18.65 versus 13.80 on Mosaic3D s2t. On SceneDepict → Sr3D, it leads with

Table 11: Cross-dataset QGS results for ScanRefer and Nr3D under query perturbations (%).

	ScanRefer → Nr3D		**Nr3D → ScanRefer**	
Method	*s2t*	*t2s*	*s2t*	*t2s*
CrossOver	11.49	10.78	11.06	12.33
+ Tent	11.49	10.77	11.05	12.22
+ READ	11.50	10.89	11.05	12.45
+ SAR	11.48	10.75	11.03	12.21
+ EATA	11.74	10.94	10.95	12.51
+ TCR	12.06	**11.06**	11.52	12.64
+ HAT-VTR	11.96	10.91	13.03	12.71
+ Ours	**13.27**	10.95	**13.56**	**13.16**
Mosaic3D	5.04	2.88	5.15	3.83
+ Tent	5.03	2.38	5.14	3.61
+ READ	5.06	3.03	5.15	3.79
+ SAR	4.96	1.98	5.04	2.32
+ EATA	4.74	2.35	5.02	3.23
+ TCR	5.31	3.08	5.35	3.90
+ HAT-VTR	6.27	3.10	6.32	3.74
+ Ours	**7.02**	**3.62**	**6.46**	**4.34**

Table 12: Cross-dataset QGS results for ScanRefer and Sr3D under query perturbations (%).

	ScanRefer → Sr3D		**Sr3D → ScanRefer**	
Method	*s2t*	*t2s*	*s2t*	*t2s*
CrossOver	8.09	9.20	9.34	10.98
+ Tent	8.11	9.19	9.31	11.04
+ READ	8.09	9.31	9.25	11.09
+ SAR	8.11	9.18	9.35	11.00
+ EATA	8.10	9.35	9.11	11.67
+ TCR	8.07	9.41	10.35	11.36
+ HAT-VTR	8.07	9.00	10.57	11.36
+ Ours	**8.36**	**9.47**	**11.53**	**12.18**
Mosaic3D	2.41	2.25	3.00	3.23
+ Tent	2.38	1.73	2.89	2.17
+ READ	2.40	**2.37**	3.12	3.46
+ SAR	2.44	1.26	3.00	3.24
+ EATA	2.39	1.42	2.84	1.90
+ TCR	2.41	2.31	2.97	3.22
+ HAT-VTR	2.39	1.75	5.92	3.37
+ Ours	**3.17**	2.16	**7.09**	**4.28**

Mosaic3D, while HAT-VTR leads in both CrossOver tasks. These results show that the benefit of adaptation depends on the source–target direction and backbone rather than being uniform across transfers.

D THE USE OF LARGE LANGUAGE MODELS

Large language models were used solely as writing assistants to improve the clarity and readability of this manuscript, including sentence rephrasing and grammatical correction. They were not used to generate the core research ideas, experimental results, or analyses. The authors take full responsibility for the content of this paper.

Table 13: Cross-dataset QGS results for Nr3D and SceneDepict under query perturbations (%).

	Nr3D → SceneDepict		SceneDepict → Nr3D	
Method	*s2t*	*t2s*	*s2t*	*t2s*
CrossOver	14.41	23.06	16.39	17.84
+ Tent	14.41	22.92	16.34	17.87
+ READ	14.45	23.41	16.50	17.96
+ SAR	14.43	22.89	16.31	17.88
+ EATA	14.37	24.19	17.26	**18.26**
+ TCR	16.13	23.96	18.02	18.10
+ HAT-VTR	18.83	24.89	19.70	18.17
+ Ours	**23.24**	**27.58**	**21.04**	17.92
Mosaic3D	13.70	10.43	4.33	2.70
+ Tent	13.74	8.18	4.26	2.07
+ READ	13.71	11.09	4.42	3.01
+ SAR	13.73	5.17	4.19	1.50
+ EATA	13.73	8.01	3.97	1.74
+ TCR	14.40	11.79	4.84	3.08
+ HAT-VTR	17.73	12.26	5.74	3.09
+ Ours	**22.33**	**14.86**	**5.90**	**3.53**

Table 14: Cross-dataset QGS results for Nr3D and Sr3D under query perturbations (%).

	Nr3D → Sr3D		Sr3D → Nr3D	
Method	*s2t*	*t2s*	*s2t*	*t2s*
CrossOver	8.24	11.29	9.41	13.08
+ Tent	8.22	11.32	9.45	13.20
+ READ	8.27	11.28	9.46	13.19
+ SAR	8.21	**11.34**	9.41	13.20
+ EATA	8.21	11.29	9.61	13.62
+ TCR	8.79	11.30	10.09	13.41
+ HAT-VTR	8.47	11.28	11.28	13.09
+ Ours	**9.52**	10.83	**12.75**	**13.81**
Mosaic3D	2.32	2.08	2.67	2.28
+ Tent	2.32	2.21	2.63	1.48
+ READ	2.30	1.96	2.65	2.27
+ SAR	2.35	1.52	2.67	2.28
+ EATA	2.37	1.67	2.64	1.55
+ TCR	2.36	2.11	2.79	2.26
+ HAT-VTR	3.26	2.05	**4.24**	2.09
+ Ours	**3.37**	**2.44**	4.13	**2.58**

Table 15: Cross-dataset QGS results for SceneDepict and Sr3D under query perturbations (%).

	SceneDepict → Sr3D		Sr3D → SceneDepict	
Method	*s2t*	*t2s*	*s2t*	*t2s*
CrossOver	9.04	10.03	11.65	14.44
+ Tent	9.04	10.04	11.56	14.05
+ READ	9.00	10.04	11.72	14.91
+ SAR	9.03	10.04	11.57	13.95
+ EATA	9.06	10.10	11.83	17.61
+ TCR	9.46	10.14	12.82	15.97
+ HAT-VTR	9.50	9.86	13.30	17.71
+ Ours	**10.43**	**10.20**	**18.51**	**24.53**
Mosaic3D	2.14	1.91	5.85	6.50
+ Tent	2.27	1.51	5.88	3.36
+ READ	2.24	1.99	6.00	8.43
+ SAR	2.17	0.96	5.83	5.73
+ EATA	2.39	1.26	3.70	3.23
+ TCR	2.30	1.96	6.40	7.67
+ HAT-VTR	2.46	1.90	13.80	7.85
+ Ours	**2.67**	**2.12**	**18.65**	**11.05**

Table 16: Comparisons on t2s Recall@1 (%) on the SceneDepict dataset under text perturbations.

	Character-Level					Word-Level					Sentence-Level					
Method	OCR	CI	CR	CS	CD	SR	RI	RS	RD	IP	Formal	Casual	Passive	Active	Backtrans	**Avg.**
CrossOver	61.40	47.25	49.38	52.52	53.18	44.35	54.18	68.51	64.95	69.11	59.66	65.67	52.68	53.56	70.77	57.81
+ Tent	63.04	50.77	52.04	55.76	55.16	46.38	57.90	70.09	66.27	70.27	60.00	66.18	52.71	53.38	71.63	59.44
+ READ	63.22	51.79	53.21	56.23	55.91	47.96	58.31	70.35	66.89	70.71	60.27	66.41	53.50	53.57	71.87	60.01
+ SAR	62.21	48.34	50.28	53.62	54.13	45.29	55.42	69.35	65.52	69.45	59.79	65.84	52.68	53.56	70.99	58.43
+ EATA	64.16	52.82	54.68	56.50	56.57	49.53	62.35	70.82	67.19	70.82	60.68	66.82	53.02	54.16	72.20	60.82
+ TCR	63.30	51.86	53.29	56.24	56.44	48.27	59.83	70.41	66.82	70.55	60.51	66.72	53.48	54.30	72.07	60.27
+ HAT-VTR	58.62	47.40	49.56	51.29	51.21	45.15	55.84	65.09	61.90	65.43	56.07	62.20	49.29	50.70	66.99	55.78
+ Ours	**65.73**	**54.48**	**56.06**	**58.79**	**58.03**	**50.24**	**63.06**	**72.60**	**68.33**	**72.50**	**61.67**	**67.90**	**54.45**	**55.12**	**73.59**	**62.17**
Mosaic3D	20.61	15.72	16.43	18.81	19.11	18.58	19.16	21.83	22.33	22.96	20.85	22.81	19.12	18.84	23.86	20.07
+ Tent	19.65	13.40	13.94	17.06	18.07	16.55	16.41	19.53	20.64	21.56	20.43	22.38	18.37	17.91	23.62	18.64
+ READ	20.87	16.55	17.22	19.42	19.23	19.23	19.99	22.24	22.38	23.04	20.71	22.77	19.29	18.94	24.10	20.40
+ SAR	18.13	10.14	11.13	15.19	15.42	13.09	13.60	17.94	18.38	20.14	20.17	22.11	18.34	17.63	22.98	16.96
+ EATA	18.20	10.96	12.30	16.21	16.17	14.38	12.68	17.89	19.43	19.32	19.76	21.83	18.31	17.83	22.40	17.18
+ TCR	20.82	16.33	17.25	19.33	19.63	19.40	20.11	22.57	22.87	22.99	21.02	22.79	19.25	18.88	24.06	20.49
+ HAT-VTR	19.28	15.84	15.87	17.48	18.26	17.49	18.77	21.12	21.18	21.45	19.19	21.05	17.48	16.74	21.86	18.87
+ Ours	**23.53**	**19.16**	**19.59**	**21.80**	**22.13**	**22.27**	**23.32**	**25.28**	**25.40**	**25.48**	**22.78**	**24.62**	**20.96**	**20.72**	**26.16**	**22.88**

Table 17: Comparisons on t2s Recall@1 (%) on the Nr3D dataset under text perturbations.

	Character-Level					Word-Level					Sentence-Level					
Method	OCR	CI	CR	CS	CD	SR	RI	RS	RD	IP	Formal	Casual	Passive	Active	Backtrans	**Avg.**
CrossOver	15.49	7.89	8.44	12.38	11.27	19.73	21.98	20.47	27.47	23.83	26.86	26.30	28.66	34.07	29.67	20.97
+ Tent	13.06	4.89	6.22	9.89	8.49	18.56	20.44	19.42	26.07	21.78	26.42	25.09	27.30	34.22	29.46	19.42
+ READ	16.83	9.20	9.40	13.03	11.85	20.67	22.79	20.77	28.62	25.43	28.65	27.54	30.63	36.73	31.29	22.23
+ SAR	15.29	6.76	7.67	10.98	10.77	19.16	20.78	19.77	26.12	22.94	26.24	25.85	28.79	33.83	29.35	20.29
+ EATA	11.35	3.41	6.05	8.34	8.01	13.71	17.47	13.93	22.38	21.15	24.31	20.98	26.12	33.10	27.52	17.19
+ TCR	16.97	9.76	9.90	13.93	12.19	21.15	23.70	21.68	29.46	25.85	29.19	28.59	31.39	36.77	32.15	22.85
+ HAT-VTR	16.00	9.22	9.40	13.21	11.42	19.23	21.78	20.50	27.35	23.95	27.23	25.82	28.89	33.49	29.73	21.15
+ Ours	**18.26**	**10.81**	**10.34**	**14.69**	**13.13**	**23.10**	**25.12**	**23.31**	**31.72**	**28.39**	**31.39**	**30.78**	**33.78**	**39.42**	**34.65**	**24.59**
Mosaic3D	7.14	5.99	5.95	7.17	6.40	7.97	8.32	8.39	8.41	8.61	9.12	8.59	8.68	8.98	9.33	7.94
+ Tent	7.10	6.19	5.92	6.83	6.61	7.89	8.21	8.25	8.52	8.31	8.58	8.55	8.93	9.05	9.30	7.88
+ READ	7.07	5.92	5.83	7.18	6.46	7.77	8.32	8.29	8.32	8.49	9.03	8.55	8.78	8.81	9.13	7.86
+ SAR	7.14	6.05	5.89	7.17	6.43	7.97	8.35	8.35	8.44	8.65	9.13	8.61	8.71	9.00	9.39	7.95
+ EATA	7.18	6.12	6.03	7.11	6.56	7.97	8.36	8.36	8.58	8.71	9.19	8.63	8.75	9.05	9.35	8.00
+ TCR	7.25	6.22	6.05	7.01	6.66	7.82	8.31	8.25	8.61	8.42	8.88	8.52	8.83	8.92	9.25	7.93
+ HAT-VTR	7.17	6.27	6.27	6.93	6.80	8.24	8.90	8.51	8.88	9.29	9.13	9.10	9.09	9.64	9.52	8.25
+ Ours	**7.87**	**7.20**	**6.74**	**7.67**	**7.50**	**8.95**	**9.64**	**9.83**	**10.04**	**10.09**	**10.48**	**10.27**	**10.30**	**10.44**	**10.51**	**9.17**

Table 18: Comparisons on t2s Recall@1 (%) on the Sr3D dataset under text perturbations.

	Character-Level					Word-Level					Sentence-Level					
Method	OCR	CI	CR	CS	CD	SR	RI	RS	RD	IP	Formal	Casual	Passive	Active	Backtrans	**Avg.**
CrossOver	6.32	6.19	6.23	6.93	7.23	10.57	10.50	10.48	11.82	10.20	10.13	11.37	12.05	12.49	12.05	9.64
+ Tent	6.97	5.76	6.20	6.64	7.06	10.04	10.30	10.48	11.75	10.31	9.90	11.61	11.83	12.33	12.12	9.55
+ READ	7.04	6.44	6.67	7.24	7.41	**11.28**	10.98	10.20	12.23	10.81	11.39	11.86	12.25	12.72	12.40	10.06
+ SAR	6.79	6.46	6.50	7.14	7.61	10.74	10.92	10.70	12.08	10.61	10.87	11.75	12.56	12.93	12.60	10.02
+ EATA	5.56	5.39	5.18	6.66	6.37	9.64	9.35	9.90	11.35	10.24	9.70	10.91	11.74	11.95	11.68	9.04
+ TCR	7.33	**6.97**	**6.73**	**7.77**	**7.85**	10.94	**11.48**	**11.10**	**12.56**	**11.18**	11.72	**12.57**	**12.89**	**13.37**	12.83	**10.49**
+ HAT-VTR	6.61	6.32	6.37	6.79	7.40	9.97	10.11	9.35	11.01	9.63	10.95	11.04	11.82	11.78	11.32	9.36
+ Ours	**7.45**	6.90	**6.73**	7.61	7.64	10.81	11.11	11.08	12.49	10.87	**11.76**	12.46	**12.89**	13.20	**12.90**	10.39
Mosaic3D	**1.81**	1.83	1.79	**2.09**	1.78	2.43	2.57	2.43	2.50	2.28	2.20	2.45	2.43	2.57	2.32	2.23
+ Tent	1.15	1.18	1.21	1.42	1.12	1.93	1.85	1.62	1.86	1.66	1.72	1.81	1.76	1.83	1.78	1.60
+ READ	1.71	1.81	1.79	1.85	1.79	2.55	2.36	2.30	2.36	2.28	2.23	2.36	2.38	2.32	2.25	2.16
+ SAR	**1.81**	1.83	1.79	**2.09**	1.78	2.43	2.55	2.25	2.45	2.29	2.28	2.32	2.42	2.56	2.30	2.21
+ EATA	1.79	1.42	1.51	2.01	1.71	1.98	2.42	2.18	2.23	1.82	2.11	2.23	2.19	2.19	2.01	1.99
+ TCR	1.79	1.91	1.86	2.02	1.78	2.45	2.48	2.40	2.50	2.39	2.22	2.40	2.29	2.42	2.36	2.22
+ HAT-VTR	1.61	1.69	1.61	1.78	1.58	2.28	1.99	2.16	2.19	1.99	2.08	2.09	2.22	2.22	2.09	1.97
+ Ours	**1.81**	**1.99**	**1.96**	1.93	**2.03**	**2.73**	**2.63**	**2.53**	**2.82**	**2.76**	**3.00**	**2.82**	**2.90**	**2.79**	**2.76**	**2.50**

Table 19: Comparisons of s2t R@1 on SceneDepict at the highest corruption severity.

	Micro-Level				Meso-Level				Macro-Level						
Method	GN	OI	UN	IN	IE	OC	LD	OCS	BM	FL	RD	TSm	TSh	RXM	**Avg.**
CrossOver	73.90	44.68	55.74	33.62	75.60	74.18	76.03	62.84	54.89	56.17	75.89	75.74	75.89	75.74	65.06
+ Tent	73.62	44.82	55.60	33.90	75.74	74.18	76.03	63.26	54.47	55.89	75.89	75.74	75.89	75.74	65.05
+ READ	73.62	44.96	55.18	33.05	75.60	74.18	76.03	62.84	54.61	55.46	75.74	75.89	76.03	75.89	64.93
+ SAR	73.90	44.68	55.89	33.62	75.60	74.04	76.03	62.98	54.75	56.17	75.89	75.74	75.89	75.74	65.07
+ EATA	74.04	44.82	55.60	33.90	75.74	74.04	75.89	62.84	53.90	56.03	75.74	75.74	75.89	75.74	64.99
+ TCR	75.74	57.87	70.50	51.21	76.03	75.04	76.45	66.52	58.30	60.28	76.03	76.31	76.45	76.31	69.50
+ HAT-VTR	77.30	63.97	69.79	62.41	77.30	76.17	77.59	70.21	60.43	64.96	77.87	77.02	77.87	77.02	72.14
+ Ours	**80.43**	**72.91**	**79.72**	**75.60**	**80.00**	**79.01**	**79.01**	**73.90**	**65.82**	**71.77**	**79.86**	**78.72**	**78.72**	**78.72**	**76.73**
Mosaic3D	21.56	26.67	9.08	23.69	28.51	26.95	28.65	20.57	15.18	25.39	28.79	29.36	7.38	28.37	22.87
+ Tent	21.42	26.67	9.08	23.55	28.51	27.09	27.94	20.43	15.60	25.53	28.79	29.08	7.38	28.51	22.83
+ READ	21.70	26.52	9.22	23.83	28.51	27.09	28.51	20.71	15.89	25.53	28.65	29.50	7.38	28.79	22.99
+ SAR	21.42	27.23	8.79	23.12	27.94	26.10	27.66	19.57	15.18	24.96	27.94	28.23	7.38	27.66	22.37
+ EATA	21.42	26.24	9.36	23.55	26.95	25.82	27.38	19.57	15.46	24.54	26.95	27.80	7.38	27.09	22.11
+ TCR	21.84	28.37	11.49	24.82	29.65	27.66	30.35	22.70	16.45	27.52	30.35	30.07	7.66	29.22	24.15
+ HAT-VTR	28.79	36.31	15.18	31.35	39.29	36.74	39.01	30.50	23.97	31.91	39.72	40.00	11.63	38.44	31.63
+ Ours	**34.33**	**42.55**	**20.43**	**36.60**	**43.97**	**44.40**	**44.82**	**36.60**	**27.09**	**38.87**	**43.40**	**45.96**	**18.30**	**43.83**	**37.22**

Table 20: Comparisons on s2t R@1 on the Nr3D dataset with M3SP 3D scene perturbations.

	Micro-Level				Meso-Level				Macro-Level						
Method	GN	OI	UN	IN	IE	OC	LD	OCS	BM	FL	RD	TSm	TSh	RXM	**Avg.**
CrossOver	41.54	25.89	31.58	15.79	41.54	40.54	42.96	33.43	28.02	29.16	43.24	43.10	42.96	42.96	35.91
+ Tent	41.39	26.03	31.72	15.93	42.96	40.54	43.24	33.29	28.17	29.30	44.67	43.81	43.81	43.67	36.32
+ READ	41.96	26.32	31.29	15.08	43.10	39.83	42.67	33.57	27.74	29.02	43.53	43.53	43.67	43.67	36.07
+ SAR	41.82	25.60	31.58	15.79	41.82	39.97	42.67	33.29	28.02	29.30	44.24	43.39	43.39	43.39	36.02
+ EATA	42.11	26.03	31.58	16.50	41.96	40.26	42.53	33.00	28.02	28.88	44.24	42.67	42.82	42.67	35.95
+ TCR	45.80	32.01	39.26	28.17	44.38	44.52	43.95	34.99	34.14	32.72	46.51	45.23	45.23	45.09	40.14
+ HAT-VTR	49.08	37.84	47.65	37.98	50.21	49.08	49.36	39.12	38.83	34.85	50.21	50.36	50.78	50.36	45.41
+ Ours	**51.07**	**47.94**	**55.62**	**52.20**	**56.61**	**54.34**	**52.20**	**43.24**	**43.24**	**42.11**	**55.05**	**54.05**	**54.91**	**54.20**	**51.20**
Mosaic3D	9.67	13.66	3.84	10.24	14.65	13.80	14.08	9.25	7.54	12.52	13.80	14.51	2.84	14.37	11.05
+ Tent	9.39	13.51	3.84	10.38	14.65	13.80	13.66	9.39	7.82	12.38	14.37	14.37	2.84	14.22	11.04
+ READ	9.82	13.66	3.84	10.24	14.65	13.51	14.37	9.53	7.68	12.66	14.08	14.37	2.56	14.37	11.10
+ SAR	9.39	13.94	3.84	10.24	14.65	13.80	13.66	9.10	7.68	12.38	13.94	14.22	2.84	14.51	11.01
+ EATA	9.39	13.94	3.98	10.24	14.22	13.80	13.66	9.39	6.97	12.38	14.22	14.65	2.70	13.80	10.95
+ TCR	9.25	14.51	3.98	10.95	14.08	13.66	13.66	10.53	7.54	12.66	14.08	14.37	2.99	14.65	11.21
+ HAT-VTR	12.23	17.07	4.55	13.66	16.64	15.50	16.64	13.37	8.68	15.36	16.93	17.07	4.69	17.50	13.56
+ Ours	**13.80**	**19.35**	**6.40**	**15.93**	**19.20**	**19.35**	**19.35**	**15.65**	**9.53**	**17.92**	**18.49**	**19.49**	**5.26**	**19.35**	**15.65**

Table 21: Comparisons on s2t R@1 on the Sr3D dataset with M3SP 3D scene perturbations.

	Micro-Level				Meso-Level				Macro-Level						
Method	GN	OI	UN	IN	IE	OC	LD	OCS	BM	FL	RD	TSm	TSh	RXM	**Avg.**
CrossOver	7.11	3.98	2.99	1.28	6.83	6.54	6.26	5.69	6.54	5.69	5.97	6.26	5.69	5.83	5.48
+ Tent	6.97	3.98	3.13	1.28	6.69	6.54	6.40	5.69	6.54	5.69	5.97	5.97	5.83	5.83	5.47
+ READ	7.25	3.98	2.99	1.28	6.83	6.83	6.40	5.55	6.54	5.69	5.97	6.12	5.83	5.97	5.52
+ SAR	7.11	4.13	3.13	1.28	6.83	6.69	6.26	5.69	6.54	5.55	5.97	6.26	5.83	5.83	5.51
+ EATA	7.54	3.98	2.99	1.28	6.97	6.97	6.83	5.69	6.54	5.41	6.12	6.26	5.97	5.97	5.61
+ TCR	8.53	5.41	4.84	3.41	9.25	8.25	8.68	7.54	7.97	7.68	9.10	8.53	8.11	8.68	7.57
+ HAT-VTR	10.38	**8.11**	6.12	**6.97**	10.81	9.25	9.96	10.10	7.97	9.10	10.10	9.39	9.39	8.53	9.01
+ Ours	**12.66**	7.54	**9.25**	5.69	**12.09**	**13.23**	**12.38**	**12.38**	**9.53**	**11.10**	**12.52**	**12.52**	**13.37**	**12.52**	**11.20**
Mosaic3D	2.28	1.85	1.42	1.99	1.99	1.56	1.71	2.13	2.28	2.84	1.56	1.85	1.71	1.99	1.94
+ Tent	2.28	1.85	1.42	2.13	2.13	1.99	2.13	2.28	2.13	2.42	1.99	2.28	1.71	1.99	2.05
+ READ	1.99	1.71	1.56	2.13	1.71	1.56	1.71	2.42	2.28	2.70	1.56	1.99	1.56	1.99	1.92
+ SAR	2.28	1.85	1.42	1.99	1.99	1.56	1.71	2.13	2.28	2.84	1.56	1.85	1.71	1.99	1.94
+ EATA	2.28	1.85	1.42	1.99	1.99	1.56	1.71	2.13	2.28	2.84	1.56	1.85	1.71	1.99	1.94
+ TCR	2.28	1.56	1.99	1.99	1.56	1.56	1.56	1.99	2.28	2.56	1.56	1.56	1.28	1.99	1.84
+ HAT-VTR	**2.42**	3.13	**2.42**	3.41	1.99	3.13	2.42	3.56	1.85	2.70	2.42	2.28	**2.42**	3.27	2.67
+ Ours	**2.42**	**4.13**	1.56	**3.70**	**4.84**	**3.56**	**3.84**	**5.12**	**3.13**	**3.27**	**4.41**	**4.41**	**2.42**	**4.84**	**3.69**