3D OPEN-MODALITY LEARNING WITH INCOMPLETE POINT-CLOUD SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal pretraining enriches point-cloud representations with image and text supervision, but commonly relies on paired point-cloud data. When point clouds are unavailable for some training objects, their image–text information cannot directly supervise the point encoder used at deployment. We study 3D Open-Modality learning, where point clouds cover only a subset of training objects and inference uses real point clouds alone. We propose *Two-stage Point Supervision Transfer (TPST)*, which separates privileged supervision transfer from calibration of the final point readout. First, *Text-Conditioned Point Replay (TCPR)* constructs image-conditioned proxy tokens, modulates them with same-object text, and calibrates them against available real-point representations. These proxies carry supervision from objects without point clouds through the retained point encoder, while real points provide geometric grounding. Second, *Geometry-Preserved Shared Readout (GPSR)* adds a shared residual readout with a bounded scalar gate while freezing the Stage-1 point base and proxy module. The complete point-only deployment keeps two frozen readouts: an independently trained geometry branch (B0) and the calibrated branch. A frozen source-only selector chooses the branch on Top-1 disagreement, and the calibrated scores may rerank only the B0 Top-5 support, so the final Top-5 set is unchanged. At 60% point-cloud availability, *TPST* improves zero-shot Top-1 accuracy over matched B0 by 0.65 pp on ModelNet40, 2.07 pp on ScanObjectNN OBJ_ONLY, and 0.87 pp on PB_T50_RS. We further evaluate representation quality through frozen-encoder few-shot linear probing and controlled ablations. Code, configurations, and evaluation scripts will be released with the paper.

1 INTRODUCTION

Point-cloud representation learning provides a foundation for object recognition and 3D scene understanding. Recent methods enrich these representations by transferring visual and linguistic knowledge through cross-modal pretraining (Afham et al., 2022; Xue et al., 2023; Liu et al., 2023; Qi et al., 2023; Zhou et al., 2025a). By aligning 3D features with image and text representations, these approaches expand the semantic supervision available to point-cloud models. However, obtaining paired 3D data remains a practical constraint. Point clouds require scanning or reconstruction and can be affected by occlusion, noise, and reconstruction failures (Dai et al., 2017; Uy et al., 2019). Consequently, an image–text corpus may contain many objects without usable point clouds, leaving part of its semantic supervision inaccessible to conventional point-based pretraining.

Related learning paradigms address different aspects of this data asymmetry. Learning Using Privileged Information (LUPI) uses auxiliary information during training but typically assumes an observed primary modality for each training example (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). Missing-modality methods use masking, completion, or adaptive fusion to support learning and prediction with incomplete inputs (Neverova et al., 2016; Ma et al., 2021; Wang et al., 2023a; b). These strategies offer useful tools, yet they do not by themselves determine how an object without a point cloud should supervise the point encoder retained for deployment. We study this problem in a 3D Open-Modality setting: each training object provides an image and text, point clouds are observed for only a fixed subset of objects, and deployment takes a real point cloud as its sole sample-level input. Point availability remains fixed throughout training, so an object without a point cloud cannot reveal one in a later epoch.

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.

Learning under this asymmetry presents three challenges. **First**, supervision must reach the point path. Objects without point clouds supply neither real point tokens nor direct geometric targets, and improvements to auxiliary image–text branches need not benefit the deployed encoder. **Second**, semantic calibration must preserve the geometry learned from real points. Images and text provide category and appearance cues, but these cues do not fully specify local shape and spatial structure. Allowing proxy objectives to continue modifying the point base during readout calibration can therefore perturb its geometric representation. **Third**, training must use the readout retained at inference. Transferring semantics into a shared base does not ensure that proxy and real-point features train the same final readout. If these features are optimized through different readouts, improvements on proxies may not carry over to the point-only inference interface. We organize TPST around three design goals: route point-missing supervision through the retained point path, freeze the deposited base while calibrating the terminal readout, and use one shared readout for real and replay features.

To address these challenges, we propose *Two-stage Point Supervision Transfer (TPST)*, a framework for learning point representations under incomplete point-cloud supervision. Its deposit–then–extract protocol uses *Text-Conditioned Point Replay (TCPR)* to transfer privileged supervision and *Geometry-Preserved Shared Readout (GPSR)* to calibrate the deployment readout. The two stages address the three challenges through the following design choices. **First**, to supervise the retained point path, *TCPR* constructs image-conditioned proxy tokens through learnable queries and modulates them with the corresponding text through bounded channel-wise scaling and shifting. For objects with observed point clouds, detached real-point tokens and features calibrate the proxies at both levels, while ordinary point objectives train the base with real geometry. For objects without point clouds, detached proxies pass through the retained point-side encoder, allowing point-to-text supervision and image-logit anchoring to update this encoder without accessing missing point clouds or reconstructing XYZ coordinates. **Second**, to preserve the learned geometric base during semantic calibration, *GPSR* freezes both the point base and the proxy module after the first stage. It calibrates the frozen features through a zero-initialized residual branch with bounded scaling, so semantic updates modify the terminal readout while leaving the base unchanged. **Third**, to align training with the deployment interface, we apply the same *GPSR* readout to real-point and replay features throughout the second stage. Real points provide the deployment-consistent anchor, while replay features extend semantic coverage. Only this shared readout is optimized in the second stage. At deployment, a frozen selector allows the calibrated readout to rerank only B0's Top-5 classes; both branches receive only real points at inference.

We evaluate *TPST* on ModelNet40 (Wu et al., 2015) and ScanObjectNN (Uy et al., 2019) through point-only zero-shot classification and frozen-encoder few-shot linear probing. Experiments use fixed object-level point availability of 80% and 60%, with complete point-cloud coverage as a reference. At 60% availability, *TPST* improves zero-shot Top-1 accuracy over matched B0 by 0.65 pp on ModelNet40, 2.07 pp on ScanObjectNN OBJ_ONLY, and 0.87 pp on PB_T50_RS. Few-shot evaluation trains only a linear classifier on frozen features; at 60% availability, TPST reaches $94.09 \pm 2.80\%$ on ModelNet40 and $80.28 \pm 4.93\%$ on ScanObjectNN OBJ_ONLY in the 5-way 20-shot setting. Controls with cross-category text donors and stopped replay-to-readout gradients examine whether the gains track correctly associated semantic supervision and the declared replay gradient path. In zero-shot evaluation, fixed class-name embeddings define the classifier; no sample-specific image or text is provided at inference.

Our main contributions are:

- We study 3D Open-Modality learning under fixed training-time point-cloud scarcity and point-only deployment, focusing on how point-missing objects can contribute supervision to the retained point representation while preserving point-only deployment.
- We introduce *TCPR*, which calibrates conditional proxy tokens using available real points and transfers image–text supervision from point-missing objects through the retained point encoder.
- We develop *GPSR* within a deposit–then–extract training protocol, freezing the learned base and proxy module while calibrating a shared readout that is directly reused for point-only inference.

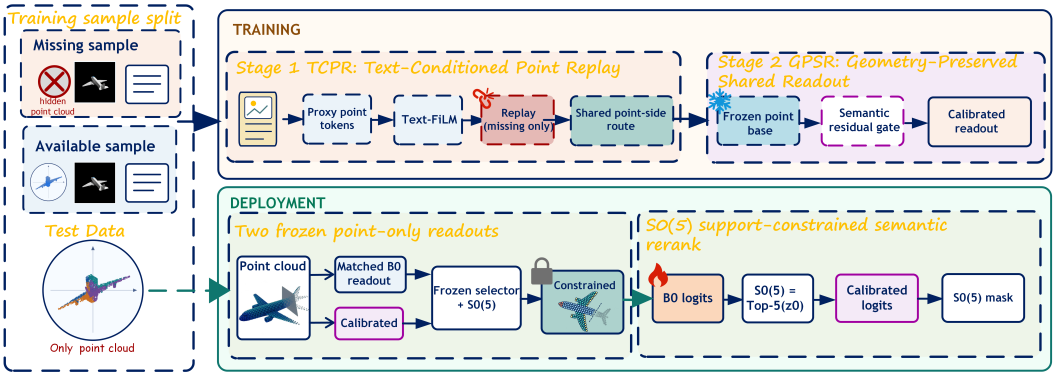

Figure 1: **TPST for point-only deployment** (§3). Top: TCPR routes real points and text-conditioned proxies through the point path, then GPSR freezes the deposited base and trains a shared terminal readout. Bottom: two frozen point-only readouts, a source-calibrated selector, and a rerank constrained to B0's Top-5 support.

2 Related Work

LUPI and Privileged Cross-Modal Learning. LUPI uses auxiliary training modalities that are removed at test time (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016; Hinton et al., 2015; Hoffman et al., 2016; Garcia et al., 2018). Generic applications include PKDOT for facial-expression recognition and LUGPI for text classification with generated visual privileged information (Aslam et al., 2024; Menadil et al., 2024); neither studies point-cloud representation learning. In 3D, image–language alignment (Zhang et al., 2022b; Zhu et al., 2023; Xue et al., 2023; 2024; Liu et al., 2023; Girdhar et al., 2023; Lei et al., 2024; Zhou et al., 2025a) and cross-modal transfer or pretraining (Qi et al., 2023; Zhou et al., 2025b; Wang et al., 2025; Sarkar et al., 2025; Jaritz et al., 2020; Chen et al., 2023) generally observe the target point cloud for every training object. Unpaired or multimodal-to-unimodal distillation can relax cohort pairing (Dou et al., 2020; Chen et al., 2022), while 2DPASS still requires LiDAR for each transferred sample (Xu et al., 2022). TPST instead transfers same-object image–text supervision from point-missing objects through a frozen-base, shared point-only readout constrained by B0.

Open-Modality Learning. Open-Modality learning allows modality sets to vary across samples or phases. Related terminology appears in modality-openness representations (Zhang & Ji, 2022), while named multimodal studies mainly address audio-visual data. OpenSR studies asymmetric cross-modal zero-shot transfer, whereas LMIM masks complete-modality training data for test-time changes (Cheng et al., 2023; Yang et al., 2025). We instead isolate fixed object-level point absence during training and real-point-only inference: TCPR transfers missing-object supervision, and GPSR trains the shared deployment readout.

Missing Modality Learning. Missing-modality learning targets inference-time modality loss (Wu et al., 2026; Neverova et al., 2016; Havaei et al., 2016). Methods impute inputs (Yang et al., 2024; Dai et al., 2025; Ke et al., 2025; Wang et al., 2026), complete or distill features (Wang et al., 2020b; Zhang et al., 2024a; Ma et al., 2021; Wang et al., 2020a; 2023a; b; Xu et al., 2026), or adapt fusion and routing (Ge et al., 2023; Yao et al., 2024; Malitesta et al., 2024; Zhan et al., 2024; Zeng et al., 2022; Zhang et al., 2023). Our missingness is instead a training-data limitation with fixed point-only deployment. Compared with the proxy-token method of Reza et al. (2025), TPST adds same-object text conditioning, available-point calibration, missing-object replay through the retained route, and a frozen-base two-readout rule.

3 Method

3.1 Problem Formulation

Training provides privileged image–text supervision, whereas deployment uses a point-only readout. "Missing" means object-level absence of the point cloud throughout pre-training, not partial point

dropout; TPST deposits this supervision with TCPR and extracts a deployment-consistent readout with GPSR after the base is frozen (Fig. 1).

Let $\mathcal{D} = \{o_n\}_{n=1}^{N}$ denote the pre-training corpus. Each object provides image I_n, text T_n, and an availability indicator a_n for point cloud P_n; write $\mathcal{D}_{\mathrm{a}} = \{o_n : a_n = 1\}$ and $\mathcal{D}_{\mathrm{m}} = \{o_n : a_n = 0\}$. Under stable-ID availability, a_n is fixed across epochs, and P_n is inaccessible to the forward pass, losses, model selection, and tuning when $a_n = 0$.

3.2 Stage 1: Text-Conditioned Point Replay

Partition parameters into the retained point route $\theta_r = \theta_{\mathrm{tok}} \cup \theta_{\mathrm{route}} \cup \theta_{\mathrm{proj}} \cup \theta_{\mathrm{scale}} \cup \theta_{\mathrm{adapt}}$ and training-only relay parameters ϕ. The point tokenizer, point-side transformer/experts, point projection, point logit scale, and deployment-persistent point adapters are in θ_r; the frozen CLIP image/text encoders are not trainable. Stage 1 updates θ_r and ϕ, whereas Stage 2 freezes θ_r, ϕ and updates only ψ. TCPR uses real points for available objects and paired image–text proxies for point-missing objects. For image tokens $H_n^I \in \mathbb{R}^{L_I \times d}$, frozen text feature $h_n^T \in \mathbb{R}^d$, and $K = 64$ learnable queries $Q \in \mathbb{R}^{K \times d}$, cross-attention yields the proxy tokens

$$\widetilde{P}_n = \mathrm{CA}(Q, H_n^I) \in \mathbb{R}^{K \times d}. \tag{1}$$

The dimensions of $\widetilde{P}_n$ match the real point-tokenizer output, so it is a token-space surrogate rather than an XYZ reconstruction. A zero-initialized text FiLM layer maps h_n^T to $(\gamma_n, \beta_n) = W_{\mathrm{film}} h_n^T + b_{\mathrm{film}}$, followed by bounded channel-wise modulation; its strength is one and $\alpha = \beta_{\max} = 0.5$.

$$P_n^{\mathrm{film}} = \widetilde{P}_n \odot \big(1 + \alpha \tanh(\gamma_n)\big) + \beta_{\max} \tanh(\beta_n), \tag{2}$$

For $a_n = 1$, let $Z_n^{p,\mathrm{tok}}$ and $z_n^{p,\mathrm{feat}}$ be the real point token sequence and routed semantic feature. The proxy tokens are first FiLM-modulated, then concatenated with the point CLS token at the shared interface and sent directly through the retained route, without a second XYZ-tokenizer pass. The CLS token is excluded from token calibration. We detach both real targets and calibrate the proxy before and after the route with normalized cosine distance d:

$$\mathcal{L}_{\mathrm{cal}} = \lambda_{\mathrm{tok}} d\big(P_n^{\mathrm{film}}, \mathrm{sg}\big(Z_n^{p,\mathrm{tok}}\big)\big) + \lambda_{\mathrm{feat}} d\big(R(P_n^{\mathrm{film}}; \theta_r), \mathrm{sg}\big(z_n^{p,\mathrm{feat}}\big)\big), \tag{3}$$

where $R(\cdot; \theta_r)$ is the retained point route and sg denotes stop-gradient. For $a_n = 0$, the conditioned proxy is detached before routing, so the missing-object losses update θ_r but not ϕ. For paired feature matrices $A, B \in \mathbb{R}^{b \times d}$ after distributed gathering, define

$$\mathcal{C}(A, B; s) = \tfrac{1}{2}\big[\mathrm{CE}\big(sAB^{\mathsf{T}}, \mathrm{diag}\big) + \mathrm{CE}\big(sBA^{\mathsf{T}}, \mathrm{diag}\big)\big]. \tag{4}$$

The diagonal entries are paired positives and all other gathered entries are negatives. Let $Z^r, Z^I, Z^T \in \mathbb{R}^{b \times d}$ be routed replay, frozen image, and frozen text features for missing rows, and let $L^r = s_p Z^r (Z^T)^{\mathsf{T}}$ and $L^I = s_I Z^I (Z^T)^{\mathsf{T}}$. The missing-object objective is

$$\mathcal{L}_{\mathrm{miss}} = \lambda_{\mathrm{pt}} \mathcal{C}(Z^r, Z^T; s_p) + \lambda_{\mathrm{anc}} d_{\mathrm{logit}}\big(L^r, \mathrm{sg}(L^I)\big). \tag{5}$$

Here d_{logit} is the standardized-logit KL image anchor. For every row u, $\widehat{u} = (u - \mu(u))/(\sigma(u) + 10^{-7})$, where the statistics are taken over the paired text entries, and $d_{\mathrm{logit}}(U, V) = \mathrm{KL}(\mathrm{softmax}(\widehat{V}/T) \,\|\, \mathrm{softmax}(\widehat{U}/T))$ is averaged over rows with $T = 1$. In the Stage-1 recipe used for the main results, $\lambda_{\mathrm{tok}} = \lambda_{\mathrm{feat}} = 0.1$, $\lambda_{\mathrm{pt}} = 0$, and $\lambda_{\mathrm{anc}} = 0.25$; the replay contrastive term is activated only in Stage 2. The first-stage parameters are obtained by jointly optimizing θ_r and ϕ,

$$\theta_r^*, \phi^* = \arg\min_{\theta_r, \phi} \mathcal{L}_{\mathrm{deposit}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{deposit}}$ combines the available-point objectives and $\mathcal{L}_{\mathrm{cal}}$ on $\mathcal{D}_{\mathrm{a}}$ with $\mathcal{L}_{\mathrm{miss}}$ on $\mathcal{D}_{\mathrm{m}}$:

$$\mathcal{L}_{\mathrm{deposit}} = \mathbb{E}_{o_n \in \mathcal{D}_{\mathrm{a}}}\big(\mathcal{L}_{\mathrm{avail}} + \mathcal{L}_{\mathrm{cal}}\big) + \mathbb{E}_{o_n \in \mathcal{D}_{\mathrm{m}}}\big(\mathcal{L}_{\mathrm{miss}}\big), \tag{7}$$

Here $\mathcal{L}_{\mathrm{avail}}$ denotes the ordinary image–text and real-point symmetric contrastive objectives on $\mathcal{D}_{\mathrm{a}}$. The two expectations are normalized separately over $\mathcal{D}_{\mathrm{a}}$ and $\mathcal{D}_{\mathrm{m}}$; an empty subset contributes zero. Thus available rows train the ordinary real-point path and proxy calibration, whereas missing rows train the retained route only through detached replay tokens.

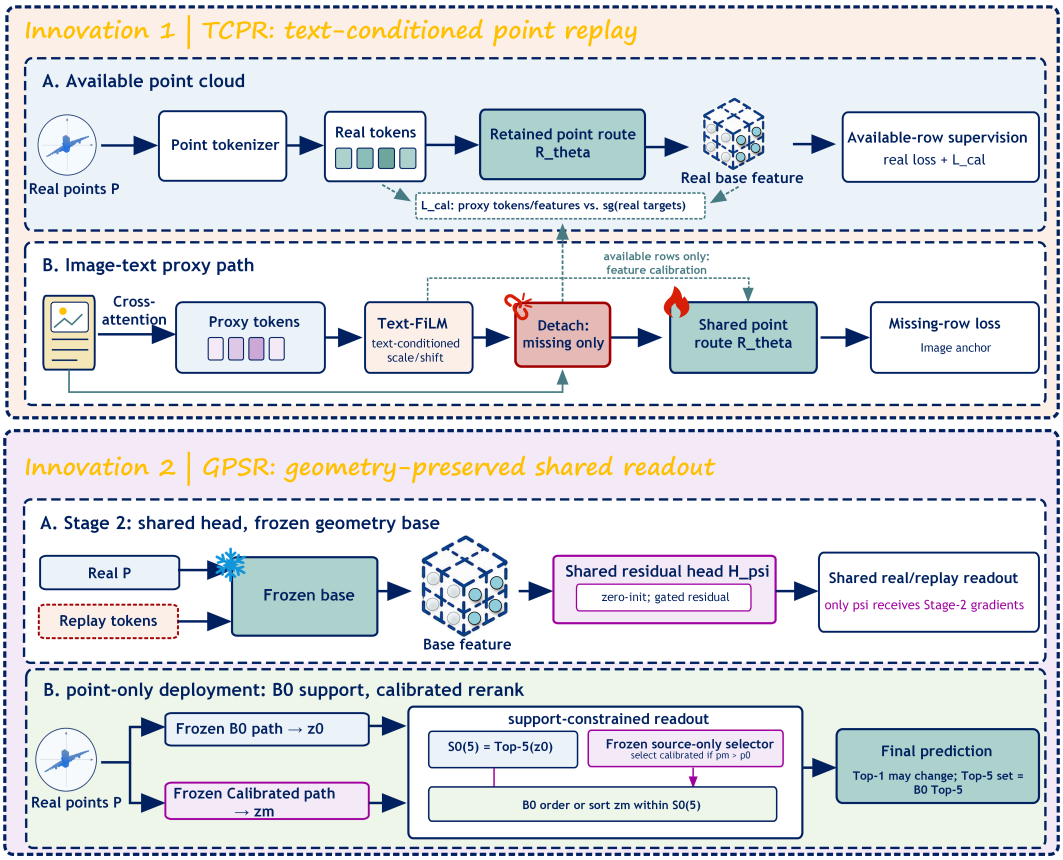

Figure 2: **TCPR and GPSR details** (§3). TCPR calibrates image–text proxies with available points and routes them through the retained point path. GPSR freezes the deposited base and trains one shared residual readout; deployment uses B0-defined Top-5 support for constrained reranking.

3.3 Stage 2: Geometry-Preserved Shared Readout

GPSR freezes the deposited point base and proxy module and provides one terminal interface for both paths. Here "geometry-preserved" means that Stage 2 does not update the deposited base; it does not assert invariant output-feature geometry. For real points, $F_{\theta_r^*} = R_{\theta_r^*} \circ \mathrm{Tok}$; for replay tokens, $F_{\theta_r^*} = R_{\theta_r^*}$ after the shared CLS/token interface. Let $z_{\mathrm{base}} = F_{\theta_r^*}(X)$ be the frozen representation. GPSR then applies the shared residual readout

$$z_{\mathrm{out}} = H_\psi(z_{\mathrm{base}}) = \mathrm{Norm}\big(z_{\mathrm{base}} + \alpha_\psi(z_{\mathrm{base}}) r_\psi(z_{\mathrm{base}})\big), \tag{8}$$

where $u = \mathrm{GELU}(W_1\,\mathrm{LN}(z_{\mathrm{base}}))$, $r_\psi = W_r u + b_r$, and the scalar gate is $\alpha_\psi = \mathrm{sigmoid}(w_g^\top u + b_g) \in [0, 1]$. The hidden width is 128. The residual-output and gate layers are zero-initialized, giving $r_\psi = 0$ and $\alpha_\psi = 0.5$ initially; Stage 2 updates only ψ. The same H_ψ serves both paths.

3.4 Two-Stage Optimization and Deployment

TPST runs TCPR and GPSR sequentially. Stage 2 starts from the fixed final Stage-1 checkpoint (epoch 99), freezes θ_r^*, ϕ^*, and uses a new head-only AdamW optimizer. For an available-row batch, let $Z^p = F_{\theta_r^*}(P)$, $Z^{r,a} = F_{\theta_r^*}(P^{\mathrm{film}})$, and $Z^{\mathrm{sem}} = \mathrm{Norm}[\frac{1}{2}\mathrm{Norm}(Z^I) + \frac{1}{2}\mathrm{Norm}(Z^T)]$. Define $L^p = s_p H_\psi(Z^p)(Z^T)^\top$ and $L^I = s_I Z^I (Z^T)^\top$. For a missing-row batch, use $Z^{r,m}$ and the

corresponding $L^r = s_p H_\psi(Z^{r,m})(Z^T)^\mathsf{T}$. The terms that depend on ψ are

$$\mathcal{L}_{\text{point}} = \mathcal{C}(H_\psi(Z^p), Z^T; s_p) + \mathcal{C}(Z^I, H_\psi(Z^p); s_I) + d_{\text{logit}}(L^p, \text{sg}(L^I)) + \underbrace{0.1\, d(H_\psi(Z^p), \text{sg}(Z^{\text{sem}}))}_{\mathcal{L}_{\text{sem}}} + 0.1\, d(H_\psi(Z^{r,a}), \text{sg}(H_\psi(Z^p))), \tag{9}$$

$$\mathcal{L}_{\text{replay}} = \rho_e \left[0.5\, \mathcal{C}(H_\psi(Z^{r,m}), Z^T; s_p) + 0.25\, d_{\text{logit}}(L^r, \text{sg}(L^I))\right]. \tag{10}$$

All logits above are batch-by-paired-text matrices; normalization in d_{logit} remains row-wise. The first objective is averaged over $\mathcal{D}_{\text{a}}$ and includes available-row replay-to-real feature calibration; the second is averaged over $\mathcal{D}_{\text{m}}$ only, so available objects are not reused as missing-object replay rows. With zero-indexed epoch e, $\rho_e = \min(1, \max(0, (e-2)/2))$. Image–text and token-calibration terms are constant with respect to ψ after freezing and are omitted from this effective head objective. The optional ranking-preservation loss has nominal weight zero. Thus

$$\psi^* = \arg\min_\psi \left[\mathbb{E}_{\mathcal{D}_{\text{a}}} \mathcal{L}_{\text{point}} + \mathbb{E}_{\mathcal{D}_{\text{m}}} \mathcal{L}_{\text{replay}}\right]. \tag{11}$$

A source-only selection over $N \in \{8, 16, 32\}$ chooses the eight-epoch Stage-2 budget before target evaluation; the selection and restart protocol is detailed in the SM.

Deployment keeps two independently trained frozen checkpoints. B0 is the matched point-only geometry baseline trained from the common initialization with the same 100-epoch budget, stable-ID mask, and evaluator but without TCPR; the calibrated checkpoint is TCPR followed by GPSR. For logits $z_0, z_m \in \mathbb{R}^C$, let $c_0 = \arg\max_c z_0[c]$, $c_m = \arg\max_c z_m[c]$, and $S_0 = \text{Top5}(z_0)$, breaking ties by class ID. Frozen source-only logistic models estimate branch-correctness scores p_0, p_m from margins, entropies, and disagreement; the fit protocol is given in the SM.

Let p_0, p_m be the two correctness scores and $g = \mathbf{1}[c_0 \neq c_m \ \wedge \ p_m > p_0]$; equality falls back to B0. If $g = 0$, the output keeps B0's ranking. If $g = 1$, the calibrated logits rank only S_0, giving $c_{m|S_0} = \arg\max_{c \in S_0} z_m[c]$ and an ordered Top-5 obtained by sorting $\{z_m[c] : c \in S_0\}$. Hence the final Top-5 set is exactly S_0 for every sample. The scores model raw branch correctness, not correctness of $c_{m|S_0}$, so the selector is a frozen branch-choice heuristic rather than a calibrated probability for the constrained final event. Both branches receive the same real point cloud; fixed class-name prototypes are the only text-derived model constants.

4 Experiments

4.1 Experimental Setup

Pre-training data and point availability. We use the point-image-text triplets of ULIP-ShapeNet (Xue et al., 2023) as the controlled pre-training corpus. For the Open-Modality protocol, every object is assigned a stable sample identifier before training. We then construct class-stratified point-availability masks at 80% and 60%, together with a 100% full-point reference. The assignment is persistent across all epochs: an available object always retains its point cloud, whereas a point-absent object exposes only its image and text. Hidden point clouds are never accessed by training, proxy calibration, model selection, or hyperparameter tuning. Source category metadata is used only to construct the stratified masks and the cross-category donor control; it is not added as a pre-training target. Matched Stage-1 comparisons use the same mask, common initialization, epoch count, and realized real-point exposure, while method-specific Stage-2, B0, and selector costs are reported separately. In the baseline controls, "Drop point-absent" removes missing rows from the training pool, whereas "Naive modality masking" retains the full fixed schedule and sends a registered zero point tensor through its point forward; neither is claimed to match TPST's optimizer-step count. This fixed sample-level protocol differs from stochastic modality dropout such as ModDrop (Neverova et al., 2016), in which an object may reveal its point cloud in a later epoch.

Downstream datasets. We evaluate on ModelNet40 (Wu et al., 2015), which contains 12,311 CAD objects from 40 categories, and ScanObjectNN (Uy et al., 2019), which contains real scanned objects from 15 categories. ModelNet40 evaluates transfer to clean synthetic geometry, whereas ScanObjectNN tests whether the frozen point representation remains useful under background clutter, occlusion, missing regions, and sensor noise. For ScanObjectNN zero-shot evaluation we report all three standard variants, OBJ_ONLY, OBJ_BG, and PB_T50_RS. The few-shot table uses OBJ_ONLY; the supplementary material provides its protocol and reproducibility details.

Table 1: **Published zero-shot references and our 100% point-availability comparison.** External results follow their original or published re-evaluation protocols; differences in pre-training data, backbones, point sampling, prompts, and ScanObjectNN splits are detailed in the SM. Our B0 and TPST rows form a matched comparison.

Method	3D train	MN40 Top-1↑	MN40 Top-5↑	Scan split	Scan Top-1↑	Protocol
PointCLIP [CVPR22]	No	20.18	–	Not specified	15.38	Original
PointCLIP V2 [ICCV23]	No	64.22	–	OBJ_ONLY	50.09	Original
CLIP2Point [ICCV23]	Yes	49.50	81.20	OBJ_ONLY	25.50	Published re-eval.
ReCon [ICML23]	Yes	61.20	78.10	OBJ_ONLY	42.30	Published re-eval.
ULIP (PointBERT) [CVPR23]	Yes	60.40	84.00	Not specified	48.50	Original
OpenShape (PointBERT) [NeurIPS23]	Yes	84.40	98.00	OBJ_ONLY	52.20	Original
ULIP-2 (PointBERT) [CVPR24]	Yes	84.70	97.10	Not reported	–	Original
UNIALIGN [CVPR25]	Yes	55.90	84.30	581-sample test	42.50	Original
Matched B0	Yes	53.61	77.07	OBJ_ONLY	37.87	Ours, 100%
TPST (ours)	Yes	54.30	77.07	OBJ_ONLY	39.50	Ours, 100%

Point-only evaluation protocols. For zero-shot classification, we freeze the model and compare normalized point features with fixed class-name prototypes from the text encoder (Radford et al., 2021; Zhang et al., 2022b; Xue et al., 2023; Liu et al., 2023; Zhou et al., 2025a). No sample-level image, text, proxy, teacher, or privileged branch is used. We report ModelNet40 Top-1/Top-5 and Top-1 on all three ScanObjectNN variants.

For few-shot probing, the pre-trained point stack remains frozen and only a new linear classifier is trained in four episode settings. Its input is the normalized 512-dimensional post-projection feature, including H_ψ for the calibrated checkpoint. The selector and constrained reranking are not used; split, point-count, and manifest details are provided in the SM.

Implementation and Reproducibility. Following UNIALIGN (Zhou et al., 2025a), we use a CLIP-initialized ViT-B/16 with a point tokenizer, LoRA-enhanced modality-aware experts (Hu et al., 2022), and a projection head. Stage 1 uses AdamW for 100 epochs, learning rate 1×10^{-4}, and weight decay 0.05; GPSR adds eight head-only epochs, while B0 is a separate 100-epoch run. No target label or hidden point selects these settings. Matched Stage-1 runs share initialization, augmentation, and evaluation seeds; other run details are in the SM.

4.2 Point-Only Zero-Shot Classification

Protocol. We evaluate frozen point encoders on ModelNet40 and all three ScanObjectNN variants using point-only input and fixed class prototypes. ModelNet40 reports Top-1/Top-5 and ScanObjectNN reports Top-1. Because pre-training captions may contain ScanObjectNN category names, the latter measures open-vocabulary transfer rather than unseen-label generalization.

Table 1 places our 100% matched pair beside published zero-shot results from PointCLIP, PointCLIP V2, CLIP2Point, ReCon, ULIP, OpenShape, ULIP-2, and UNIALIGN (Zhang et al., 2022b; Zhu et al., 2023; Huang et al., 2023; Qi et al., 2023; Xue et al., 2023; Liu et al., 2023; Xue et al., 2024; Zhou et al., 2025a). These external rows use different training data and evaluation details and therefore provide context rather than a controlled ranking.

Results. The controlled comparison in Table 1 is the matched 100% B0–TPST pair; availability-controlled results are reported separately below. At 100%, where no objects lack points, TPST raises Top-1 on all four benchmarks; this setting does not measure missing-row replay. Its final Top-5 is exactly B0's $S_0(5)$ and is preserved by construction: only ordering within that support can change.

Table 2: **Zero-shot results under fixed point-cloud availability.** All rows use the same stable-ID protocol and strict point-only evaluation. TPST denotes the frozen selector with B0-defined $S_0(5)$ constrained reranking. Top-5 is inherited from the matched B0 support by construction.

Variant	Avail. (%)	MN40 Top-1↑	MN40 Top-5↑	OBJ_ONLY Top-1↑	OBJ_BG Top-1↑	PB_T50_RS Top-1↑
Matched B0	100	53.61	**77.07**	37.87	32.87	20.54
TPST	100	**54.30**	**77.07**	**39.50**	**34.40**	**21.60**
Matched B0	80	50.65	74.64	39.41	33.73	22.21
TPST	80	**51.46**	**74.64**	**39.93**	**34.42**	**22.45**
Matched B0	60	51.74	**75.41**	34.77	32.70	20.61
Drop point-absent objects	60	50.57	74.35	**36.83**	32.19	**21.96**
Naive modality masking	60	49.47	74.47	**36.83**	31.67	21.86
TPST	60	**52.39**	**75.41**	**36.83**	**33.39**	21.48

Table 3: **Few-shot linear** probing on ModelNet40 and ScanObjectNN OBJ_ONLY with a frozen point encoder (§4.3). Results are mean accuracy ± standard deviation over 10 fixed episodes with identical manifests. The TPST row uses the frozen calibrated checkpoint feature; the selector and $S_0(5)$ are not used.

Method	Avail. (%)	MN40 5w10s↑	MN40 5w20s↑	MN40 10w10s↑	MN40 10w20s↑	Scan 5w10s↑	Scan 5w20s↑	Scan 10w10s↑	Scan 10w20s↑
Matched B0	60	94.16 ±2.99	94.64 ±2.76	89.89 ±3.98	91.42 ±3.01	78.93 ±4.92	80.35 ±5.61	66.94 ±3.97	70.87 ±2.75
TPST	60	93.08 ±3.30	94.09 ±2.80	89.56 ±3.57	90.99 ±2.86	79.10 ±6.02	80.28 ±4.93	66.46 ±3.11	69.84 ±2.08

Table 2 compares fixed 100%, 80%, and 60% availability levels under the same stable-ID protocol. At 60%, TPST improves matched B0 Top-1 by 0.65 pp on ModelNet40 and by 2.07, 0.69, and 0.87 pp on OBJ_ONLY, OBJ_BG, and PB_T50_RS, respectively. At 80%, its corresponding gains are 0.81, 0.52, 0.69, and 0.24 pp; these are descriptive point estimates.

4.3 Few-Shot Linear Probing

Protocol. Few-shot linear probing measures class separability without updating the point representation. All checkpoints share episode manifests, support/query identities, augmentations, linear-head optimization, and trial seeds, isolating pre-training quality from full fine-tuning effects.

The zero-shot gains do not translate into consistent improvements in frozen-feature few-shot probing. The 60% TPST feature is slightly higher in one ScanObjectNN setting but lower in the remaining settings; without paired statistical analysis, we treat these as descriptive differences.

4.4 Ablation Studies

Setup. We conduct all main ablations on ModelNet40 with 60% fixed point availability. A reported comparison must use the same stable-ID mask, common initialization, Stage-1 epoch budget, applicable Stage-2 budget, and strict point-only evaluator. Short 8-epoch diagnostics are not mixed with the 100-epoch endpoints reported in Table 4.

Contribution of Each Component. Table 4 uses four implementation-level switches. "Calib." is token/feature proxy calibration on available rows against stopped real-point targets. "Replay" routes detached missing-object proxies through the retained route; in the main Stage-1 recipe this contributes only the image-logit anchor with weight 0.25. $\mathcal{L}_{\text{sem}}$ is the available-row residual-readout alignment in the Stage-2 objective. "R→ H_ψ" indicates whether missing-row replay losses update the shared terminal readout in Stage 2. A removal disables only that named term or gradient path; ordinary real-point objectives and the shared deployment readout remain active.

Semantic-Association and Gradient-Path Controls. Table 5 reports descriptive controls rather than uniquely identified causal effects. For each point-absent receiver, the cross-category control replaces the complete auxiliary text record with a one-to-one donor from a different source category.

Table 4: **Components and deployment readout on ModelNet40 at 60% point availability.** Model rows report raw zero-shot scores after the stated training stages; the final row additionally applies the frozen selector and B0-defined $S_0(5)$ constrained readout. Stage-1 rows use 100 epochs and applicable Stage-2 rows use eight head-only epochs.

Variant	Calib.	Replay	$\mathcal{L}_{\text{sem}}$	R$\rightarrow H_\psi$	Top-1↑	Top-5↑
Matched B0 (raw)	–	–	–	–	51.74	75.41
Proxy calibration only (raw)	✓	–	–	–	51.46	73.42
Full P1+P2 (raw)	✓	✓	✓	✓	51.09	73.26
Full w/o replay-to-readout gradient (raw)	✓	✓	✓	–	51.58	73.42
TPST + selector/$S_0(5)$ (deployed)	✓	✓	✓	✓	**52.39**	**75.41**

Receiver identity and image, point availability, and the auxiliary-text marginal are preserved; no donor image or point cloud is loaded. Donor text supplies both the FiLM condition and all text-derived targets for that missing row. The mask, seed, and epochs are fixed, but the donor checkpoint uses proxy-calibration weight 0 instead of 0.1; batch size and optimizer-step counts may also differ. The gradient-path control keeps the shared deployment readout but makes replay losses consume the frozen pre-head replay feature, stopping their gradient to H_ψ; the available path is unchanged.

Table 5: **Semantic-association and gradient-path controls** on ModelNet40 at 60% point availability. All method rows use the final frozen selector and $S_0(5)$ readout. Delta Top-1 is relative to the correct-association row. The donor and starred availability rows are descriptive controls.

Privileged signal	Receiver path	Top-1↑	Top-5↑	Δ Top-1↑
Correct associated text	Retained point path	52.39	75.41	0.00
Cross-category text donor	Retained point path	52.19	75.41	-0.20
Correct associated text	Replay-to-readout gradient stopped	52.51	75.41	$+0.12$
No point-absent objects	Removed during training	50.57	74.35	-1.82^*

5 Conclusion

Existing 3D multimodal pretraining assumes complete point–image–text triplets and cannot use image–text objects without points. We study fixed point-cloud availability with point-only deployment. *TPST* follows a deposit–then–extract protocol: *TCPR* routes image–text-conditioned proxies through the retained point path, and *GPSR* freezes this base while training a shared terminal readout. At 60% availability, deployed TPST improves matched B0 Top-1 by 0.65 pp on ModelNet40 and 2.07 pp on ScanObjectNN OBJ_ONLY while preserving B0's $S_0(5)$. The zero-shot gains do not consistently transfer to few-shot probing. Deployment uses two frozen point-only readouts, a source-calibrated selector, and within-support reranking; conclusions are limited to the evaluated fixed-availability protocols.

AI use statement

We used AI assistants for language polishing and structured review of technical exposition, including checks of definitions, notation, claim scope, and reproducibility wording. The authors verified all technical statements, citations, protocols, and reported results.

Reproducibility Statement

Section 3 specifies TPST and deployment, and Section 4.1 gives the experimental protocol. SM Sections A–B record implementation settings, fixed availability masks, data/model resources, and evaluation details. We will release the corresponding code and reproducibility artifacts.

REFERENCES

Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *CVPR*, pp. 9892–9902, 2022.

Zixiang Ai, Zichen Liu, Yuanhang Lei, Zhenyu Cui, Xu Zou, and Jiahuan Zhou. GAPrompt: Geometry-aware point cloud prompt for 3D vision model. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 796–808, 2025.

Muhammad Haseeb Aslam, Muhammad Osama Zeeshan, Soufiane Belharbi, Marco Pedersoli, Alessandro Lameiras Koerich, Simon Bacon, and Eric Granger. Distilling privileged multimodal information for expression recognition using optimal transport. In *FG*, pp. 1–10, 2024.

Cheng Chen, Qi Dou, Yueming Jin, Quande Liu, and Pheng-Ann Heng. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Trans. Med. Imaging*, 41(3):621–632, 2022. doi: 10.1109/TMI.2021.3119385.

Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. CLIP2Scene: Towards label-efficient 3d scene understanding by CLIP. In *CVPR*, pp. 7020–7030, 2023. doi: 10.1109/CVPR52729.2023.00678.

Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. Opensr: Open-modality speech recognition via maintaining multi-modality alignment. In *ACL*, pp. 6592–6607, 2023.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017. doi: 10.1109/CVPR.2017.261.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *ICCV*, pp. 24507–24517, 2025.

Qi Dou, Quande Liu, Pheng-Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE Trans. Med. Imaging*, 39(7):2415–2425, 2020. doi: 10.1109/TMI.2019.2963882.

Yi Du, Zhipeng Zhao, Shaoshu Su, Sharath Golluri, Haoze Zheng, Runmao Yao, and Chen Wang. Superpc: A single diffusion model for point cloud completion, upsampling, denoising, and colorization. In *CVPR*, pp. 16953–16964, 2025.

Chenghao Fang, Jianqing Liang, Jiye Liang, Hangkun Wang, Kaixuan Yao, and Feilong Cao. Multimodal point cloud completion with interleaved attention enhanced transformer. In *IJCAI*, pp. 963–971, 2025.

Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *ECCV*, pp. 106–121, 2018. doi: 10.1007/978-3-030-01237-3_7.

Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *ICCV*, pp. 8721–8731, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *MICCAI*, pp. 469–477, 2016. doi: 10.1007/978-3-319-46723-8_54.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *CVPR*, pp. 826–834, 2016. doi: 10.1109/CVPR.2016.96.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W. H. Lau, Wanli Ouyang, and Wangmeng Zuo. CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. In *ICCV*, pp. 22157–22167, 2023.

Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, pp. 12602–12611, 2020. doi: 10.1109/CVPR42600.2020.01262.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *CVPR*, pp. 25864–25873, 2025.

Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. ViT-Lens: Towards omni-modal representations. In *CVPR*, pp. 26647–26657, 2024.

An Li, Zhe Zhu, and Mingqiang Wei. Genpc: Zero-shot point cloud completion via 3d generative priors. In *CVPR*, pp. 1308–1318, 2025.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. In *NeurIPS*, 2023.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI*, volume 35, pp. 2302–2310, 2021.

Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Fragkiskos D Malliaros, and Tommaso Di Noia. Dealing with missing modalities in multimodal recommendation: A feature propagation-based approach. *arXiv preprint arXiv:2403.19841*, 2024.

Rafael-Edy Menadil, Mariana-Iuliana Georgescu, and Radu Tudor Ionescu. Learning using generated privileged information by text-to-image diffusion models. In *ICPR*, pp. 423–438, 2024.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, pp. 604–621, 2022.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pp. 652–660, 2017a.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, pp. 5099–5108, 2017b.

Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3D representation learning guided by generative pretraining. In *ICML*, pp. 28223–28243, 2023.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *NeurIPS*, 2022.

Jiahui Qu, Yuanbo Yang, Wenqian Dong, and Yufei Yang. Lds2ae: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities. In *AAAI*, volume 38, pp. 14731–14739, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. Robust multimodal learning via cross-modal proxy tokens. *Trans. Mach. Learn. Res.*, 2025.

Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. CrossOver: 3D scene cross-modal alignment. In *CVPR*, pp. 8985–8994, 2025.

Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, pp. 6411–6420, 2019.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, pp. 1588–1597, 2019. doi: 10.1109/ICCV.2019.00167.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Netw.*, 22(5-6):544–557, 2009.

Hao Wang, Yanyu Qian, Pengcheng Weng, Zixuan Xia, William Dan, Yangxin Xu, and Fei Wang. COMPASS: Complete multimodal fusion via proxy tokens and shared spaces for ubiquitous sensing. *arXiv preprint arXiv:2604.02056*, 2026.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *CVPR*, pp. 15878–15887, 2023a.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *KDD*, pp. 1828–1838, 2020a.

Qianqian Wang, Huanhuan Lian, Gan Sun, Quanxue Gao, and Licheng Jiao. icmsc: Incomplete cross-modal subspace clustering. *IEEE Trans. Image Process.*, 30:305–317, 2020b.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *ICCV*, pp. 22025–22034, 2023b.

Ziyi Wang, Yanran Zhang, Jie Zhou, and Jiwen Lu. UniPre3D: Unified pre-training of 3D point cloud models with cross-modal gaussian splatting. In *CVPR*, pp. 1319–1329, 2025.

Guangshun Wei, Yuan Feng, Long Ma, Chen Wang, Yuanfeng Zhou, and Changjian Li. Pcdreamer: Point cloud completion through multi-view diffusion priors. In *CVPR*, pp. 27243–27253, 2025.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Trans. Mach. Learn. Res.*, 2026. arXiv:2409.07825.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *CVPR*, pp. 4840–4851, 2024.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pp. 1912–1920, 2015. doi: 10.1109/CVPR.2015.7298801.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Junyi Wu, Rui Xu, and Jinglin Xu. MCMoE: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In *AAAI*, volume 40, pp. 11241–11249, 2026. doi: 10.1609/aaai.v40i13.38104.

Yan Xu, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2DPASS: 2d priors assisted semantic segmentation on LiDAR point clouds. In *ECCV*, pp. 677–695, 2022. doi: 10.1007/978-3-031-19815-1_39.

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

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP-2: Towards scalable multimodal pre-training for 3d understanding. In *CVPR*, pp. 27081–27091, 2024. doi: 10.1109/CVPR52733.2024.02558.

Yanwu Yang, Hairui Chen, Zhikai Chang, Yang Xiang, Chenfei Ye, and Ting Ma. Incomplete learning of multi-modal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In *MIDL*, pp. 1006–1018. PMLR, 2024.

Zhao Yang, Rui Jiang, Xiao Fu, Wei Xi, and Jizhong Zhao. Open-modality latent modality interaction maximization for audio-visual learning. In *ICASSP*, pp. 1–5. IEEE, 2025.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *AAAI*, volume 38, pp. 16416–16424, 2024.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, pp. 19291–19300, 2022.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *EMNLP*, pp. 2924–2934, 2022.

Yaohua Zha, Yanzi Wang, Hang Guo, Jinpeng Wang, Tao Dai, Bin Chen, Zhihao Ouyang, Xue Yuerong, Ke Chen, and Shu-Tao Xia. Pma: Towards parameter-efficient point cloud understanding via point mamba adapter. In *CVPR*, pp. 16976–16986, 2025.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. In *ACL*, pp. 9637–9662, 2024.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *ICML*, pp. 41753–41769, 2023.

Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *NeurIPS*, pp. 27061–27074, 2022a.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In *CVPR*, pp. 8552–8562, 2022b.

Yiyuan Zhang and Yuqi Ji. Modality eigen-encodings are keys to open modality informative containers. In *ACM MM*, pp. 4122–4131, 2022.

Zheyu Zhang, Gang Yang, Yueyi Zhang, Huanjing Yue, Aiping Liu, Yunwei Ou, Jian Gong, and Xiaoyan Sun. Tmformer: Token merging transformer for brain tumor segmentation with missing modalities. In *AAAI*, volume 38, pp. 7414–7422, 2024a.

Zhiyuan Zhang, Licheng Yang, and Zhiyu Xiang. Risurconv: Rotation invariant surface attention-augmented convolutions for 3d point cloud classification and segmentation. In *ECCV*, pp. 93–109. Springer, 2024b.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pp. 16259–16268, 2021.

Bo Zhou, Liulei Li, Yujia Wang, Huafeng Liu, Yazhou Yao, and Wenguan Wang. Unialign: Scaling multimodal alignment within one unified model. In *CVPR*, pp. 29644–29655, 2025a.

Wei Zhou, Qian Wang, Weiwei Jin, Xinzhe Shi, and Ying He. Graph transformer for 3d point clouds classification and semantic segmentation. *Comput. Graph.*, 124:104050, 2024.

Ziyang Zhou, Pinghui Wang, Zi Liang, Haitao Bai, and Ruofei Zhang. Cross-modal 3D representation with multi-view images and point clouds. In *CVPR*, pp. 3728–3739, 2025b.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *ICCV*, pp. 2639–2650, 2023.

Supplementary Material

This supplementary material provides reproducibility details, extended related work, and a scoped discussion of limitations and social impact. Section A states the reproducibility and release scope; Section B reports the execution environments, resource identities, training configuration, and evaluation protocols; Section C presents extended related work; and Section D discusses limitations and social impact.

A Reproducibility Statement

Reproducibility statement. The main paper specifies the prediction function and effective objectives. This SM identifies the execution environments, stable-ID inputs, initialization, resource roles, and downstream protocols used by the reported runs. Run-level logs and resolved configurations are recorded during execution; the appendix below lists the fields needed to identify them but does not claim that every run log is embedded in the manuscript.

Reproducibility scope. TPST has two sequential stages. TCPR deposits image–text supervision while updating the retained point route and its training-only proxy module. GPSR then freezes both and trains only the shared terminal readout from the fixed final Stage-1 checkpoint (epoch 99). The recorded prerequisite checkpoint allows Stage 2 to be rerun without repeating Stage 1. "Geometry-preserved" means only that GPSR does not update the deposited base; it does not claim invariant output-feature geometry.

We distinguish the input contract from the TPST algorithm. Point-only compliance requires that each test sample provide only a real point cloud; fixed class-name prototypes are allowed, but sample-level image, text, proxy, teacher, or hidden-point inputs are not. A single-checkpoint point model can satisfy this input contract. Full TPST deployment is the more specific two-readout procedure in the main paper: B0 and the calibrated checkpoint are frozen, the source-only selector is frozen, and calibrated scores may reorder only B0's Top-5 support.

Traceable experimental artifacts. For each training run we record the stable-ID manifest and mask versions, initialization identifier, random seed, resolved arguments, optimizer, learning rate, batch size, accumulation, world size, epoch count, checkpoint identifier, runtime, and realized real-point exposure. For evaluation we record the point-input files, fixed class prototypes, feature readout, episode manifests, support/query identities, trial seeds, and checkpoint identifier. These execution records support internal audit; the planned release will provide the corresponding code, configurations, derived masks, manifests, commands, and machine-readable summaries.

Release and data access. We plan to release implementation code, dependency specifications, resolved configurations, stable-ID masks, few-shot manifests, and training/evaluation commands together with version identifiers for scripts, checkpoints, and dataset files. We will not redistribute raw ULIP-ShapeNet, ModelNet40, or ScanObjectNN data. Users must obtain those assets from their official sources, and released derived artifacts will retain upstream attribution and license notices.

B Experiment Setting

Software and hardware environments. Reported runs use one of the two recorded stacks in Table A1. Each result manifest binds a run to its actual machine and resolved configuration; versions are not pooled across machines.

Table A1: **Recorded computing environments.** The result manifest identifies the stack used by each run.

Machine	Software stack	Use
RTX 4090 (24 GB)	Python 3.8.20; PyTorch 2.2.2; CUDA 12.1; OpenCLIP 2.26.1; Transformers 4.41.2	Training/evaluation
2× RTX 3090 (24 GB)	Python 3.8.10; PyTorch 2.0.0; CUDA 11.7; OpenCLIP 2.32.0; Transformers 4.46.3	Training/evaluation

Training configuration. Stage 1 trains B0 and TCPR independently from the same common initialization for 100 runtime epochs with AdamW, an initial learning rate of 1×10^{-4}, and weight decay

0.05. GPSR starts from the fixed final TCPR checkpoint (epoch 99), uses a fresh head-only AdamW optimizer, and runs the selected exact budget of eight epochs, saving epoch 7. Batch size, accumulation, world size, optimizer-step count, wall time, and peak memory are run-level fields rather than one manuscript-wide constant. The source-only selector is fitted after both point checkpoints are frozen and performs no point-model update.

Token calibration and Stage-2 selection. Real and replay token sequences both have length 64. Token calibration averages index-wise cosine distances under the fixed tokenizer/query ordering; this is token-index alignment rather than spatial correspondence. Real targets include the tokenizer's point positional embeddings, whereas replay tokens receive no second positional embedding. If sequence lengths differ, each sequence is mean-pooled before computing one cosine distance. Stage-2 candidates $N \in \{8, 16, 32\}$ use a deterministic 10% source split stratified by availability and category. Selection minimizes the mean held-out point–text, image–point, and shared semantic losses and chooses the smallest N within 1% of the minimum. This selects $N = 8$; the final run restarts from the frozen Stage-1 checkpoint with a fresh optimizer and saves epoch 7.

Selector fitting. For frozen branch logits z_0, z_m, a source-only 2:1 fit/lock split trains two ℓ_2-regularized logistic models for the raw correctness events of B0 and the calibrated branch. Their shared input is $(m_0, m_m, m_0 - m_m, e_0, e_m, e_0 - e_m, \mathbf{1}[c_0 \neq c_m], m_0/(m_m + 10^{-8}))$, where m is the raw-logit Top-1 margin and e is softmax entropy. Five-fold source cross-validation minimizes Brier loss over $C_{\mathrm{LR}} \in \{0.1, 1, 10\}$; the lock split is diagnostic only, and all coefficients are frozen before target evaluation.

Datasets, models, and resource roles. The controlled pre-training corpus contains 51,637 stable ULIP-ShapeNet object IDs. CLIP/OpenCLIP supplies the frozen image and text encoders and the ViT-B/16 visual initialization. The UNIALIGN release is used as a code and architecture reference for the point tokenizer, modality-aware route, LoRA modules, and projection interface; no pretrained UNIALIGN 3D weights are loaded. Point-side parameters are trained from the common initialization identified below. Table A2 separates these roles from downstream data use.

Table A2: **Dataset and model resources.** Raw data and external weights remain subject to their upstream terms.

Asset	Source type	Role
ULIP-ShapeNet	Public dataset release	Controlled pre-training corpus
ModelNet40	Public benchmark	Zero-/few-shot evaluation
ScanObjectNN	Public benchmark	Zero-/few-shot evaluation
CLIP / OpenCLIP	Public model release	Frozen encoders; visual initialization
UNIALIGN	Public code release	Code/architecture reference; no 3D weights loaded

Stable-ID and initialization identities. The controlled pre-training manifest contains 51,637 stable object IDs. The class-stratified 80% mask contains 41,308 available and 10,329 missing objects; the 60% mask contains 30,982 available and 20,655 missing objects; and the 100% mask contains 51,637 available and zero missing objects. The manifest, masks, and common initialization are versioned in the reproducibility package rather than embedded as long identifiers in the manuscript. Category metadata is used only to stratify masks and construct the cross-category donor control, not as an added pre-training target. Hidden point clouds from missing objects are not used to initialize weights or construct caches or targets.

Downstream evaluation. Strict zero-shot evaluation freezes the point model, uses 2,048 XYZ points per sample, and compares its normalized semantic feature with fixed class-name prototypes. The selector and support-constrained rerank are used only for the full TPST zero-shot deployment. Few-shot probing instead trains a new linear classifier on the normalized 512-dimensional post-projection feature, including H_ψ for the calibrated checkpoint; neither the selector nor constrained reranking is used.

ModelNet40 few-shot episodes pool the official train and test populations, sample the requested classes, take k support objects per class, and use every remaining selected-class object as a disjoint query, with 1,024 points per object. ScanObjectNN episodes sample supports from the official training split and use every selected-class object in the official test split as a query, with 2,048 points per object. The four settings are 5-way 10-shot, 5-way 20-shot, 10-way 10-shot, and 10-way 20-shot. Seeds 0–9 determine class and support sampling; stored manifests bind class choices, support/query

Table A3: **Protocols for the published zero-shot references in Table 1.** "Original" denotes the method paper; "published re-eval." denotes evaluation of a released checkpoint reported by OpenShape. A split is left unspecified when the source paper does not name it.

Method	Backbone	Pre-training data	3D training	Scan split	MN40 / Scan source note
PointCLIP	CLIP RN50	CLIP image–text pre-training	No	Not specified	Original; Top-5 not reported
PointCLIP V2	CLIP ViT-B/32	CLIP and GPT prompting	No	OBJ_ONLY	Original, best settings
CLIP2Point	ViT-B/32	ShapeNet image–depth pairs	Yes	OBJ_ONLY (581)	Published re-eval.
ReCon	ViT-B/Transformer	ShapeNet multimodal data	Yes	OBJ_ONLY (581)	Published re-eval.
ULIP	PointBERT	ShapeNet55 multimodal data	Yes	Not specified	Original Tables 4–5
OpenShape	PointBERT	ShapeNetCore, Objaverse, 3D-FUTURE, ABO	Yes	OBJ_ONLY (581)	Original complete ensemble
ULIP-2	PointBERT	Objaverse + ShapeNet	Yes	Not reported	Original; zero-shot Scan absent
UNIALIGN	ViT-B	Complete-point multimodal training	Yes	581-sample test	Original; split name not stated
Ours	ViT-B/16	ULIP-ShapeNet, 51,637 stable IDs	Yes	OBJ_ONLY	Matched B0/TPST; strict point-only

Table A4: **Few-shot controls at 60% point availability.** Each cell reports mean accuracy ± standard deviation over 10 fixed episodes. The encoder is frozen and only a newly initialized linear classifier is trained; Scan denotes OBJ_ONLY. DROP removes point-absent rows, while NAIVE uses the registered zero-point input.

Method	Avail. (%)	MN40 5w10s↑	MN40 5w20s↑	MN40 10w10s↑	MN40 10w20s↑	Scan 5w10s↑	Scan 5w20s↑	Scan 10w10s↑	Scan 10w20s↑
DROP	60	93.81 ±3.12	94.40 ±2.82	90.03 ±3.89	91.52 ±2.89	77.35 ±4.95	79.89 ±5.47	66.14 ±2.27	69.33 ±2.67
NAIVE	60	93.21 ±3.72	94.07 ±2.88	90.02 ±3.94	91.53 ±2.92	78.69 ±4.51	81.78 ±3.13	68.06 ±2.69	72.69 ±2.41

identities, and recorded data versions and are reused across compared checkpoints. These are two explicitly different protocols rather than one shared notion of domain shift.

Published zero-shot protocol differences. The external results in Table 1 were checked against the corresponding publications (Zhang et al., 2022b; Zhu et al., 2023; Huang et al., 2023; Qi et al., 2023; Xue et al., 2023; Liu et al., 2023; Xue et al., 2024; Zhou et al., 2025a). CLIP2Point and ReCon use OpenShape's published checkpoint re-evaluation on the 581-sample OBJ_ONLY split; the other external rows use the stated original-paper result. Table A3 expands the main table's compact 3D-training indicator and records the principal protocol differences, so these rows are context rather than a common-protocol ranking.

Few-shot availability controls. Table A4 reports the two 60% controls omitted from the main few-shot table. Both use the same ModelNet40 and ScanObjectNN OBJ_ONLY episode manifests as the main comparisons.

C Extended Related Work

Modality Completion. Recent advances in modality completion have been dominated by diffusion models due to their powerful generative capabilities, albeit at the cost of significant computational resources. In the 3D vision domain, GenPC (Li et al., 2025) achieves zero-shot point cloud completion by leveraging pre-trained generative priors, while PCDreamer (Wei et al., 2025) recovers global structures from single-view partial clouds via multi-view diffusion priors. SuperPC (Du et al., 2025) unifies multiple point cloud restoration tasks—including completion, upsampling, and denoising—within a single diffusion framework. For multimodal scenarios, IAET (Fang et al., 2025) proposes

an interleaved attention mechanism to bidirectionally complement incomplete point clouds with RGB image features. In remote sensing, LDS2AE (Qu et al., 2024) handles random modality missing for classification. Nevertheless, these methods make different assumptions about missingness and validate within their stated settings; their modeling assumptions and validation scope address those settings rather than arbitrary modal loss. *TCPR* avoids explicit test-time modality completion: its image-conditioned proxy tokens are modulated by same-sample text through bounded FiLM, routed through the retained point-side components only during training, and removed at point-only inference.

3D Object Classification. 3D object classification is a standard benchmark for point-cloud representation learning. Foundational point-set architectures include PointNet, PointNet++, and KPConv (Qi et al., 2017a; b; Thomas et al., 2019). Transformer and self-supervised pretraining methods include Point Transformer, Point-BERT, Point-MAE, Point-M2AE, and PointNeXt (Zhao et al., 2021; Yu et al., 2022; Pang et al., 2022; Zhang et al., 2022a; Qian et al., 2022).

Recent progress has improved efficient backbones, geometric robustness, and parameter-efficient adaptation. PointCLIP, ULIP-2, and CLIP2Scene improve cross-modal transfer, while Point Transformer V3, GTNet, RISurConv, GAPrompt, and PMA develop stronger backbones or efficient adaptation mechanisms (Zhang et al., 2022b; Xue et al., 2024; Chen et al., 2023; Wu et al., 2024; Zhou et al., 2024; Zhang et al., 2024b; Ai et al., 2025; Zha et al., 2025). These methods generally assume complete point-cloud supervision and focus on classifier or representation design.

Our work is orthogonal to 3D classifier design. We use 3D object classification as a downstream probe of point-only representation quality, and study how privileged image-text supervision affects this representation when point-cloud training samples are incomplete. Strict zero-shot testing probes semantic alignment, while frozen-encoder few-shot linear probing measures class separability under the two stated episode protocols. These conclusions are scoped to the evaluated point backbone and protocols.

D Limitations and Social Impact

Limitations. Our method has two practical limitations. First, *TCPR* relies on the quality of paired image–text supervision and on calibration from objects with available point clouds; noisy pairs or weak proxy calibration may reduce transfer to point-missing objects. Second, our study uses fixed class-stratified availability masks on one pre-training corpus and evaluates object classification, so it does not yet cover arbitrary missingness, other 3D tasks, or all deployment conditions. Future work can test whether confidence-aware proxy filtering, broader and imbalanced availability masks, and validation on additional datasets and tasks alleviate these limitations. We will also explore adaptive proxy calibration and automatic stage-boundary selection while preserving point-only deployment.

Social Impact. By reusing widely available image–text resources when point clouds are scarce, *TPST* may reduce the cost of collecting 3D supervision and lower barriers to 3D perception research. However, image–text supervision may inherit biases from pretrained models, and replay proxies may contain semantic or geometric errors. We recommend documenting source datasets and models, filtering unreliable samples, and using *TPST* representations as supplementary supervision rather than a substitute for real-point validation in safety-critical or human-facing applications. The method uses public data, collects no new human information, and all released code and configurations will follow the licenses and usage conditions of their source assets.