GenCOPE: Syn2Real Generalized Category-Level Object Pose Estimation for Robotic Picking

Jian Liu[1], **Wei Sun**[2], **Zhenqi Dai**[2], **Hui Yang**[2], **Jian Xiao**[2], **Nicu Sebe**[3], **Na Zhao**[1,†]

[1] IMPL, SUTD [2] NERC-RVC, Hunan University [3] University of Trento

Project page: CNJianLiu/GenCOPE

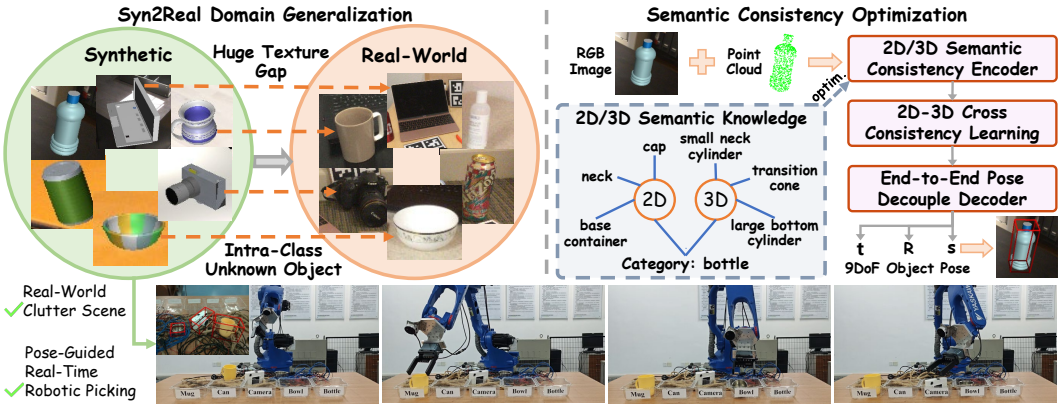

Figure 1: GenCOPE for real-world robotic picking. *Left*: synthetic-to-real (Syn2Real) domain-generalized task setting, which aims to achieve generalization between intra-class unknown objects from synthetic to real-world domains. *Right*: 2D/3D semantic consistency optimization for end-to-end pose regression, where the dashed box indicates that it is used for optimization only during training.

Abstract

Category-level object pose estimation (COPE), capable of generalizing to intra-class unknown objects, has become a core technique for robotic 3D scene understanding. However, existing COPE methods still require labor-intensive recollection of real-world training data for novel object categories, which limits their scalability in practical applications. This paper aims to achieve synthetic-to-real (Syn2Real) generalized COPE, where a model is trained solely on rendered synthetic data and directly generalized to real-world deployments. The central challenge lies in the significant domain gap between synthetic and real-world data, particularly in texture appearance. To address this, we aim to enhance domain generalization by learning domain-invariant representations that capture semantic commonalities among objects within the same category. We introduce 2D and 3D semantic consistency constraints to reduce the sensitivity of feature encoders to domain-specific features. In addition, we propose an end-to-end pose regression framework that performs 2D-3D cross consistency learning, leveraging dense cross-modality fusion to further refine pose estimation. Since simplicity and effectiveness are essential for real-world robotic deployment, our model operates exclusively on global features, yielding a highly lightweight and efficient architecture. Extensive experiments on the REAL275 and Wild6D benchmarks, as well as real-world robotic manipulation scenes, show superior Syn2Real generalization performance of our paradigm.

[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Nine-degrees-of-freedom (9DoF) object pose estimation, which aims to estimate the 3DoF translation, 3DoF rotation, and 3DoF size of an object in the camera coordinate system [1, 2, 3, 4], is a fundamental problem in robotic 3D scene understanding and manipulation [5, 6, 7, 8, 9, 10]. Early studies primarily focused on instance-level pose estimation [11, 12, 13, 14, 15], where models are restricted to objects seen during training. Alternatively, category-level object pose estimation (COPE) has emerged as a promising direction [16, 17, 18, 19, 20, 21, 22], as it enables generalization to intra-class unknown objects without relying on explicit prior information such as CAD models or reference views.

Despite the impressive generalization capability of category-level approaches across unknown instances within the same category, existing methods still require labor-intensive collection and annotation of real-world training data for each novel object category [23, 24, 25, 26]. This requirement significantly hinders scalability and limits their practicality in large-scale robotic applications [27, 28, 29, 30]. In contrast, recent advances in rendering technology have made large-scale synthetic data readily available, with accurate pose annotations obtained at negligible cost. Motivated by this observation, this paper aims to achieve Syn2Real domain-generalized COPE, referred to as GenCOPE, where models trained purely on synthetic data can be directly deployed in real world.

The central challenge in GenCOPE lies in the significant domain gap between synthetic and real-world data. Through an analysis of the rendered synthetic data used for object pose estimation [2], we observe that the most pronounced discrepancy arises from texture appearance, as illustrated in Fig. 1. However, existing methods primarily rely on feature extractors [31, 32] that emphasize fine-grained image features [33, 34, 35], which inherently encode substantial texture information while failing to explicitly capture domain-invariant properties shared between synthetic and real-world data. Consequently, the learned representations become overly sensitive to texture variations, making them suboptimal for robust Syn2Real domain-generalized pose estimation.

To address the aforementioned challenge, we argue that although intra-class unknown objects exhibit substantial texture discrepancies between synthetic and real-world domains, their 2D and 3D semantic structures remain largely consistent. For instance, at the 2D level, objects such as *bottles* typically consist of semantic parts including *a cap, a neck, and a base container*. Similarly, at the 3D level, their geometric composition can be abstracted as *a small neck cylinder, a transitional cone, and a large bottom cylinder*. Based on this observation, we posit that explicitly incorporating semantic knowledge as constraints during 2D and 3D feature encoding can encourage encoders to focus on domain-invariant semantic cues, thereby suppressing the adverse effects of domain-specific features. An overview of the proposed GenCOPE is illustrated in Fig. 1.

Given a segmented RGB image and its corresponding point cloud (obtained via depth map back-projection) as inputs, we first perform feature embedding using the proposed 2D/3D semantic consistency encoders, followed by a 2D-3D cross consistency learning module and an end-to-end pose decouple decoder. During training, semantic consistency is enforced by leveraging semantic supervision derived from a pretrained large language model [36] and the CLIP text encoder [37], which aligns the 2D and 3D feature representations in a shared semantic space. The resulting semantically consistent features are then fused through a 2D-3D cross transformer-based module to perform cross-consistency learning [38]. Finally, the fused features are concatenated and fed into a pose decouple decoder to regress the full 9DoF object pose. To prevent feature extractors from overemphasizing fine-grained texture-dependent features while maintaining computational efficiency, we adopt the most lightweight backbones based on ResNet18 [39] and PointNet [40] as the 2D and 3D encoders, respectively, and operate exclusively on global features throughout the pipeline. This design is deliberately adopted to jointly enhance domain generalization robustness while meeting the real-time requirements of practical robotic manipulation scenarios.

Overall, our main contributions are as follows: (1) We propose a simple yet effective framework for Syn2Real domain-generalized COPE, explicitly designed for real-world robotic deployment. The framework operates purely on global features from most lightweight backbones, avoiding dense local texture correspondences, making it highly efficient under resource-constrained settings. (2) We introduce semantic consistency encoders in both 2D and 3D spaces to learn semantic-invariant features shared across intra-class objects, alleviating the severe gap between synthetic and real-world domains. (3) We propose an end-to-end pose estimation pipeline based on 2D-3D cross consistency learning, featuring a cross-modal fusion network and a pose decouple decoder, enabling mutual reinforcement across modalities for more accurate and robust pose estimation.

2 Related Work

Fully-Supervised COPE Methods. NOCS [2] introduces a normalized object coordinate space formulation and establishes two widely adopted benchmarks that advance the research of COPE methods. To address substantial shape variability among intra-class objects, SPD [41] first leverages category-level CAD model libraries to extract shape priors offline and subsequently reconstructs NOCS representations through shape prior deformation for pose estimation. This paradigm has inspired a series of follow-up approaches that explicitly incorporate shape priors [42, 43, 44, 45]. Despite their effectiveness in modeling intra-class shape variability, prior-based approaches rely on manually constructed category-level CAD model libraries, which incur substantial human effort.

The limitation aforementioned has motivated the development of prior-free alternatives [46, 47, 48, 34, 21, 49, 50, 51, 52]. FS-Net [47] proposes a 3D graph convolution-based network and performs decoupled pose regression. GPV-Pose [48] enhances the learning of pose-sensitive features via geometry-guided point-wise voting. Beyond these depth-only approaches, several RGB-D methods [34, 53] jointly leverage appearance and geometric cues. SecondPose [54] introduces DINOv2 [55] to provide SE(3)-consistent semantic features for pose estimation. GCE-Pose [56] enhances COPE under partial visibility by reconstructing category-level global context via semantic shape reconstruction and fusing it with local RGB-D observations. CleanPose [52] mitigates data bias by integrating causal inference with front-door adjustment and residual-based knowledge distillation from 3D foundation models. Although these fully-supervised methods achieve competitive performance, their dependence on extensive real-world data annotation makes them difficult to scale to novel object categories.

Label-Efficient COPE Methods. To alleviate the scalability bottleneck caused by extensive real-world data annotation, recent studies have explored label-efficient COPE [57, 58, 59, 60, 33, 61, 62, 63, 64, 65]. CPPF [57] introduces a category-level point pair feature voting formulation that enables pose estimation in the wild without dense supervision. SAR-Net [58] investigates object shape alignment by directly matching the observed point cloud with shape prior, reducing the reliance on real-world training data. DPDN [33] designs a domain-generalized learning framework based on parallel deep prior deformation. UDA-COPE [61] leverages a teacher-student self-supervised learning scheme to achieve domain adaptation. Further, TTA-COPE [62] proposes a test-time adaptation strategy that first learns from labeled synthetic data and then adapts the pretrained model online to real-world data during inference. More recently, DiffusionNOCS [64] adopts diffusion models to predict dense canonical maps for pose estimation, demonstrating strong robustness to symmetry and effective Syn2Real generalization. Diff9D [65] directly formulates COPE as a diffusion-based generative process, enabling Syn2Real generalization without requiring 3D shape priors.

Although existing methods perform well, they do not explicitly model the consistent properties shared between synthetic and real-world data. In contrast, this paper analyzes the domain-invariant characteristics across synthetic and real domains and explicitly incorporates semantic knowledge as constraints during both 2D and 3D feature encoding. This design encourages encoders to focus on domain-invariant semantic cues, alleviating the adverse impact of domain-specific information.

3 Workflow of GenCOPE

Task definition: GenCOPE aims to predict the 6DoF object pose (consisting of 3DoF translation and 3DoF rotation) and 3DoF object size (axes lengths of oriented bounding box) of intra-class unknown objects in real world using the model trained only on synthetic dataset.

3.1 Overview

GenCOPE takes a single RGB-D image as input. The target object is first segmented using Mask R-CNN [66] to obtain an object-centric RGB image I [67, 68]. The corresponding segmentation mask is then applied to crop the depth map, which is back-projected to generate a point cloud P with 1024 sampled points. This preprocessing follows the standard pipeline adopted by most prior works [42, 52, 65]. Subsequently, I and P are processed by dedicated 2D and 3D encoders equipped with semantic consistency optimization (*applied only during training*) to extract global semantic consistency features $F_{C\text{-}2D}$ and $F_{C\text{-}3D}$. The resulting features are then fused by a proposed 2D-3D cross transformer network, which explicitly models cross-modal consistency. Finally, the fused 2D-3D consistency-aware global features are concatenated and fed into a decoupled decoder to

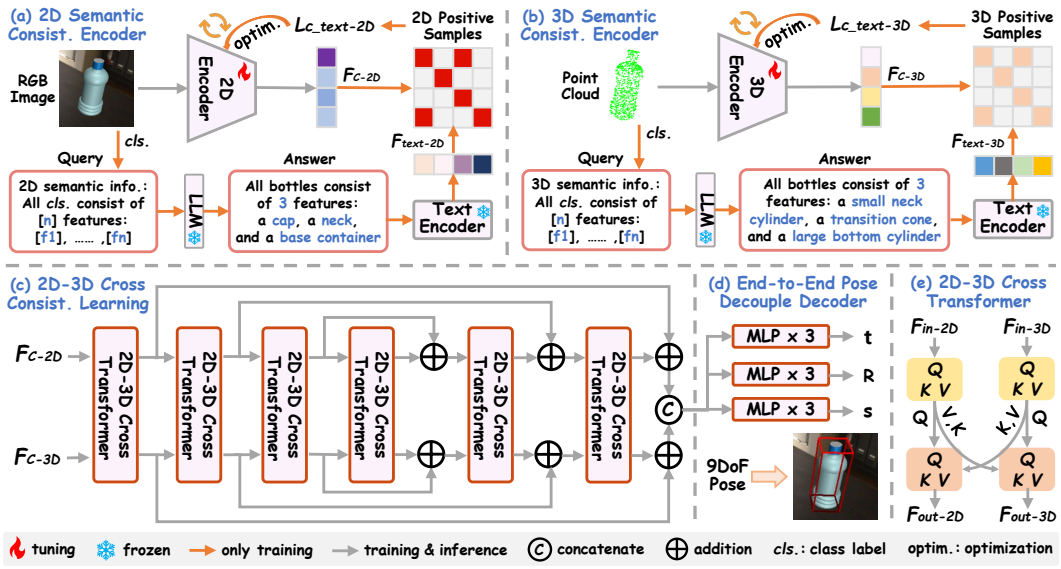

Figure 2: Overall workflow of GenCOPE. (a) and (b) illustrate the proposed 2D and 3D semantic consistency encoders, which impose training-time constraints to emphasize intra-class semantic invariances. (c) presents the proposed 2D-3D cross consistency learning module that fuses two-modal features, where the 2D-3D cross transformer architecture is detailed in (e). (d) denotes the end-to-end pose decouple decoder, which regresses the full 9DoF object pose.

regress the 3DoF translation $\mathbf{t}$, 3DoF rotation $\mathbf{R}$, and 3DoF size $\mathbf{s}$, respectively. An overview of the proposed framework is illustrated in Fig. 2.

3.2 2D and 3D Semantic Consistency Encoders

To introduce intra-class semantic commonalities into RGB image and point cloud encoders in a simple yet effective manner, we leverage semantic optimization generated by the large language model (LLM) Gemini-3 [36], as illustrated in (a) and (b) of Fig. 2. Specifically, we define a semantic query template: *Output the common 2D/3D semantic information of a class of objects enclosed in parentheses in the following format: "All cls. consist of* $[n]$ *features:* $[f_1], \ldots, [f_n]$*"*, where *cls.* denotes the segmented object class, and the number of features $[n]$ as well as the semantic descriptors $[f_1]$-$[f_n]$ are inferred by the LLM. To enhance the diversity of semantic descriptions and mitigate the impact of uncertainty, we prompt the LLM to generate **ten distinct** semantic descriptions for each object class and randomly select one description for each training sample. Since this semantic consistency optimization is applied only in the training phase, the **ground-truth class label** *cls.* for each image is directly available. For instance, when *cls.* corresponds to *bottle*, a representative 2D semantic description is *"All bottles consist of three features: a cap, a neck, and a base container"*, while the corresponding 3D semantic description is *"All bottles consist of three features: a small neck cylinder, a transition cone, and a large bottom cylinder"*. Additional LLM-generated semantic examples for different object categories are provided in the Appendix A.

The generated 2D and 3D semantic descriptions are encoded using the CLIP pretrained text encoder [37], producing semantic feature embeddings $F_{text\text{-}2D} \in \mathbb{R}^{b\times 512}$ and $F_{text\text{-}3D} \in \mathbb{R}^{b\times 512}$, where b denotes the batch size. These semantic embeddings are then employed to optimize the global features encoding processes of the 2D and 3D encoders [39, 40], respectively, resulting in semantically consistent representations $F_{C\text{-}2D} \in \mathbb{R}^{b\times 512}$ and $F_{C\text{-}3D} \in \mathbb{R}^{b\times 512}$. By explicitly encouraging the encoders to attend to intra-class semantic invariances, the proposed optimization suppresses sensitivity to discrepancies between synthetic and real-world domains, improving domain generalization ability.

We consider a batch consisting of b images and b text descriptions, and let $F^i_{C\text{-}2D}$ and $F^i_{text\text{-}2D}$ denote the 2D feature of the i-th image and the semantic feature of the i-th text, respectively. Since a training batch may contain multiple positive image-text pairs sharing the same *cls.* label, we define $\mathcal{P}(i)$ (positive sample) as the set of indices of all text samples in the batch whose *cls.* labels are identical to

that of $F^i_{C\text{-}2D}$, and $\mathcal{A}(i)$ as the index set of all text samples in $F_{text\text{-}2D}$. Based on these definitions, the 2D semantic consistency loss is formulated as:

$$\mathcal{L}_{C_text\text{-}2D} = \frac{1}{b}\sum_{i=1}^{b}\mathcal{L}_i, \tag{1}$$

where the loss for each sample i, denoted as $\mathcal{L}_i$, is defined as:

$$\mathcal{L}_i = -\frac{1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)}\log\frac{\exp\left(z_i\cdot z_p/\tau\right)}{\sum\limits_{j\in\mathcal{A}(i)}\exp\left(z_i\cdot z_j/\tau\right)}, \tag{2}$$

where z_i denote the normalized feature vector of the i-th sample from $F_{C\text{-}2D}$, z_p and z_j represent normalized feature vectors from $F_{text\text{-}2D}$, $|\mathcal{P}(i)|$ is the cardinality of the positive sample set, and τ is a scalar temperature parameter [69].

Similarly, we can compute the 3D semantic consistency loss $\mathcal{L}_{C_text\text{-}3D}$ following the same formulation, by replacing $F_{C\text{-}2D}$ and $F_{text\text{-}2D}$ with $F_{C\text{-}3D}$ and $F_{text\text{-}3D}$ in the corresponding batch. By incorporating $\mathcal{L}_{C_text\text{-}2D}$ and $\mathcal{L}_{C_text\text{-}3D}$ during training, we can enforce semantic consistency supervision of the 2D and 3D encoders, discouraging over-reliance on domain-specific information.

3.3 2D-3D Cross Consistency Learning for Pose Regression

To effectively align semantic representations across modalities and enable robust pose estimation, we introduce a 2D-3D cross consistency learning framework coupled with an end-to-end pose decouple decoder. As illustrated in (c) and (d) of Fig. 2, the proposed design jointly fuses semantically consistent features $F_{C\text{-}2D}$ and $F_{C\text{-}3D}$ via stacked cross transformers and decodes the object pose in a fully decoupled end-to-end manner. The details of some modules are provided in the Appendix C.

3.3.1 2D-3D Cross Consistency Learning

We construct a stack of $2N$ 2D-3D cross transformer blocks to progressively model cross-modal interactions at multiple semantic levels. Each cross transformer block follows the architecture shown in (e) of Fig. 2, consisting of a self-attention module and a cross-attention module. For the k-th block, self-attention is first applied independently to each modality to capture intra-modal dependencies:

$$\tilde{F}^{(k)}_{C\text{-}2D} = \mathrm{SelfAttn}\left(F^{(k-1)}_{C\text{-}2D}\right),\ \tilde{F}^{(k)}_{C\text{-}3D} = \mathrm{SelfAttn}\left(F^{(k-1)}_{C\text{-}3D}\right), \tag{3}$$

where the multi-head attention [32] is formulated as:

$$\mathrm{Attn}\left(\cdot\right) = \mathrm{Concat}\left(Softmax\left(QK^{\top}/\sqrt{d}\right)V\right), \tag{4}$$

where Q, K, and V denote the input query, key, and value, respectively. $\top$ represents matrix transpose. d is the feature dimension of each attention head. $\mathrm{Concat}(\cdot)$ denotes concatenate the features of all heads. Then, cross attention is employed to enable inter-modal feature exchange, where queries originate from the same modality and keys/values from the opposite modality:

$$\begin{aligned}\hat{F}^{(k)}_{C\text{-}2D} &= \mathrm{CrossAttn}\left(Q=\tilde{F}^{(k)}_{C\text{-}2D},\ K=\tilde{F}^{(k)}_{C\text{-}3D},\ V=\tilde{F}^{(k)}_{C\text{-}3D}\right),\\ \hat{F}^{(k)}_{C\text{-}3D} &= \mathrm{CrossAttn}\left(Q=\tilde{F}^{(k)}_{C\text{-}3D},\ K=\tilde{F}^{(k)}_{C\text{-}2D},\ V=\tilde{F}^{(k)}_{C\text{-}2D}\right).\end{aligned} \tag{5}$$

To stably aggregate semantic information across different levels and encourage consistent cross-modal alignment, we introduce a symmetric residual fusion strategy, as shown in (c) of Fig. 2. The outputs of cross transformer blocks are paired symmetrically from outer to inner layers, which explicitly combines shallow and deep semantic representations [70].

3.3.2 End-to-End Pose Decouple Decoder

The fused features from the outermost symmetric pair are concatenated to form the final joint representation F_{out}. As shown in (d) of Fig. 2, F_{out} is fed into an end-to-end pose decouple decoder

composed of multiple multi-layer perceptrons (MLPs), which independently regress translation $\mathbf{t}$, rotation $\mathbf{R}$, and size $\mathbf{s}$, yielding full 9DoF object pose:

$$\mathbf{t},\, \mathbf{R},\, \mathbf{s} \;=\; \mathrm{MLPs}(F_{\text{out}}),\; \mathrm{MLPs}(F_{\text{out}}),\; \mathrm{MLPs}(F_{\text{out}}). \tag{6}$$

We follow most previous work [34, 33, 49, 65] in using translation, rotation, and size matrices to represent the 9DoF object pose.

3.4 Training and Inference Protocol

We train the proposed GenCOPE framework in an end-to-end manner, jointly optimizing semantic consistency learning and pose regression. The overall training objective consists of two components: semantic consistency losses that supervise the feature encoders, and a pose regression loss that guides the decoupled pose decoder. Specifically, the total training loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{C_text\text{-}2D} + \mathcal{L}_{C_text\text{-}3D} + \mathcal{L}_{\text{pose}}, \tag{7}$$

where the first two items enforce semantic alignment between visual features and text-derived semantic priors within each object category, and $\mathcal{L}_{\text{pose}}$ supervises the regression of the object pose. $\mathcal{L}_{\text{pose}}$ is computed as the sum of the L2 norms of translation, rotation, and size prediction errors [34, 65], which supports the pose-decoupled decoder by supervising each pose component independently.

During training, the semantic consistency losses $\mathcal{L}_{C_text\text{-}2D}$ and $\mathcal{L}_{C_text\text{-}3D}$ are applied to constrain the 2D and 3D encoders to focus on intra-class semantic invariants, thereby reducing reliance on domain-specific cues and improving robustness to domain shifts between synthetic and real-world data. The pose loss $\mathcal{L}_{\text{pose}}$ is jointly optimized to ensure accurate pose estimation. *At inference time*, the semantic consistency supervision parts are removed. Given a single RGB-D image, the trained 2D and 3D encoders, 2D-3D cross consistency learning module, and pose decouple decoder are executed in a single forward pass to directly predict the full 9DoF object pose, **without requiring any textual input** or additional optimization. This design enables efficient and scalable deployment in real-world robotic applications.

4 Experiments

Datasets We follow previous COPE methods [59, 34, 65] to utilize the CAMERA25 [2], REAL275 [2], and Wild6D [59] datasets for evaluation. Specifically, we use the large synthetic CAMERA25 for training, which includes 1085 instances distributed in ∼250K synthetic images. During inference, we evaluate on real-world REAL275 and Wild6D datasets and self-collected robotic picking scenes.

Implementation Details The input RGB-D images are captured at a resolution of 640×480. For preprocessing, instance segmentation is first performed using Mask R-CNN [66]. The resulting object mask is applied to both the RGB and depth images, after which the RGB crop is resized to 192×192 to reduce computational overhead. The temperature parameter τ in Eq. 2 is set to 0.07 as the previous contrastive learning method [69], the feature dimension d in Eq. 4 is fixed to 32, and the batch size b is set to 48. Model optimization adopts a cyclic learning rate schedule, where the learning rate oscillates between 1×10^{-6} and 1×10^{-4} using CyclicLR [71], with a step size of 20K iterations per cycle. Our model is trained for a total of 100 cycles. All experiments on public benchmarks are conducted on a single NVIDIA RTX 3090 GPU with 24 GB memory. Real-world experiments are deployed on a physical robotic platform equipped with an NVIDIA GTX 1660 GPU with 6 GB memory and an Intel RealSense L515 RGB-D camera.

Evaluation Metrics Following common practice in prior work [42, 20, 54, 52, 65], we evaluate performance using the $n°m\text{cm}$ and 3D Intersection-over-Union (IoU_{3D}) metrics. The $n°m\text{cm}$ criterion measures pose accuracy by jointly thresholding rotation and translation errors, where a prediction is deemed correct if the rotation and translation errors fall below $n°$ and $m\text{cm}$, respectively. We report results under $5°2\text{cm}$, $5°5\text{cm}$, $10°2\text{cm}$, and $10°5\text{cm}$ settings. The IoU_{3D} metric quantifies the overlap between the estimated and ground-truth 3D bounding boxes and is primarily used to assess size estimation accuracy. A prediction is considered correct when IoU_{3D} exceeds a threshold of 75%. In addition to accuracy, we assess model deployability and efficiency by reporting the GPU memory consumption (GB) during inference and running speed (FPS).

Table 1: Comparison on the REAL275 dataset in metrics of IoU_{3D} (%) and $n°mcm$ (%). "✓" and "-" indicate with and without. "Syn" and "Real w/o Label" indicate synthetic dataset and real-world dataset without label. We use the IoU_{3D} metric of CATRE [20], which corrects NOCS [2] for size evaluation. The best results are bolded.

Method	Syn	Real w/o Label	Shape Prior	mean Average Precision (mAP)				
				$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
Segmented by Mask R-CNN [66] Pretrained on ImageNet Dataset								
SPD [41]	✓	-	✓	17.1	11.4	12.0	33.5	37.8
SGPA [42]	✓	-	✓	17.8	19.8	27.7	36.5	62.6
STG6D [72]	✓	-	✓	14.4	20.1	27.9	39.0	63.7
SAR-Net [58]	✓	-	✓	-	31.6	42.3	50.3	68.3
RePoNet [59]	✓	✓	✓	-	29.1	31.3	48.5	56.8
SSC6D+ICP [60]	✓	✓	-	-	28.6	33.4	51.8	62.9
DiffusionNOCS [64]	✓	-	-	-	-	35.0	-	66.6
CleanPose [52]	✓	-	-	39.9	39.5	49.6	51.0	66.2
Diff9D [65]	✓	-	-	41.7	35.3	43.9	54.8	70.0
GenCOPE (ours)	✓	-	-	**42.6**	**39.7**	**50.0**	**57.0**	**71.6**
Segmented by Mask R-CNN [66] Pretrained on CAMERA25 Dataset								
DPDN [33]	✓	-	✓	39.8	29.7	37.3	53.7	67.0
UDA-COPE [61]	✓	✓	-	34.4	30.5	34.9	57.0	66.1
TTA-COPE [62]	✓	-	-	39.7	30.2	35.9	**61.7**	73.2
VI-Net [53]	✓	-	-	11.3	11.7	19.9	14.4	29.1
SecondPose [54]	✓	-	-	14.7	12.7	21.2	17.6	34.2
CleanPose [52]	✓	-	-	42.6	37.9	48.7	51.8	67.8
Diff9D [65]	✓	-	-	44.1	36.5	45.2	57.7	72.2
GenCOPE (ours)	✓	-	-	**44.5**	**38.1**	**49.9**	56.4	**74.4**

4.1 Comparisons with State-of-the-Art Methods

After training on the synthetic CAMERA25 dataset [2], we first conduct quantitative comparisons between GenCOPE and state-of-the-art (SOTA) methods on the publicly available real-world REAL275 [2] and the in-the-wild Wild6D [59] benchmarks. Then, we further perform qualitative comparisons under the same settings and finally evaluate domain generalization performance in real-world robotic manipulation scenarios.

Quantitative Analysis on REAL275 and Wild6D The results are reported in Table 1 and Table 2. These two datasets represent complementary real-world challenges: REAL275 focuses on controlled real-world evaluation, while Wild6D emphasizes severe domain shift and unconstrained object appearances. For a fair comparison, for baselines with released models trained on CAMERA25, we directly evaluate them on real-world benchmarks; otherwise, we retrain using their official code on CAMERA25 only, without changing any settings, and evaluate on real-world benchmarks.

REAL275 Dataset: To comprehensively evaluate the robustness of GenCOPE under different segmentation qualities, we conduct experiments using two segmentation protocols. First, we adopt Mask R-CNN [66] pretrained on ImageNet [73] to obtain object masks. The corresponding results are shown in the upper part of Table 1. Under identical experimental settings, GenCOPE achieves the best performance across all evaluation metrics without relying on any shape priors, consistently outperforming self-supervised approaches such as RePoNet [59] and SSC6D [60], as well as domain generalization methods including DiffusionNOCS [64] and Diff9D [65]. To further ensure that no real-world training data is used throughout the entire pipeline, we employ Mask R-CNN trained solely on the synthetic CAMERA25 dataset [2] for instance segmentation. The results under this setting are summarized in the lower part of Table 1. GenCOPE again shows superior performance compared to label-efficient methods, including DPDN [33], UDA-COPE [61], TTA-COPE [62], and Diff9D [65], confirming its robustness under more challenging segmentation conditions. For a fair comparison, all methods are trained exclusively on the synthetic CAMERA25 dataset. While some methods further leverage the unlabeled REAL275 dataset [2] during training, GenCOPE does not utilize any real-world data. Moreover, when comparing against domain-specific SOTA methods with publicly available implementations such as VI-Net [53], SecondPose [54], and CleanPose [52], we

Table 2: Comparison on the Wild6D dataset. "Real" indicates training with the CAMERA25 and REAL275 datasets. "✓", "-", and "Syn" represent the same meanings as explained in Table 1.

Method	Syn	Real	Shape Prior	mean Average Precision (mAP)				
				$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
SPD [41]	✓	✓	✓	20.26	6.90	9.26	20.10	27.84
SGPA [42]	✓	✓	✓	34.46	26.28	29.18	**33.80**	39.52
GPV-Pose [48]	✓	✓	-	-	14.10	21.50	23.80	41.10
Diff9D [65]	✓	-	-	38.16	25.52	30.46	32.48	40.94
GenCOPE (ours)	✓	-	-	**38.72**	**26.66**	**33.32**	33.12	**42.74**



Figure 3: Qualitative results comparison with CleanPose [52] on REAL275 and Diff9D [65] on Wild6D. The red and green boxes denote the estimated and ground-truth results, respectively.

evaluate their models trained on CAMERA25 only, ensuring full consistency with the GenCOPE training protocol. Overall, GenCOPE outperforms all these methods on REAL275 across both IoU_{3D} and $n°mcm$ metrics, indicating a more reliable holistic pose estimation under real-world sensors.

Wild6D Dataset: The quantitative results on Wild6D dataset [59], as illustrated in Table 2, further validate the robustness of GenCOPE under extreme domain shifts and in-the-wild conditions. The Wild6D dataset contains cluttered scenes, diverse lighting conditions, and background interference, posing challenges to existing methods. GenCOPE achieves the best overall performance in terms of both IoU_{3D} and $n°mcm$ metrics, surpassing fully supervised approaches SPD [41], SGPA [42], and GPV-Pose [48], as well as Diff9D [65] under the same domain-generalized experimental setting. This proves that the cross-modal semantic consistency supervision suppresses over-dependence on domain-specific information and encourages semantic-aware feature learning. Consequently, GenCOPE maintains stable accuracy even when object appearances deviate from the training distribution.

Deployability and Efficiency Analysis: To enable practical deployment on real-world robotic platforms, model deployability and runtime efficiency are essential considerations. We therefore compare GenCOPE with CleanPose [52] and Diff9D [65] on the same computing platform, in terms of GPU memory consumption at inference and inference speed, as summarized in Table 3. GenCOPE requires the least computational resources while achieving the fastest

Table 3: Analysis of model deployability and efficiency on the same computing platform. Infer. Memory indicates GPU memory usage at inference, and Speed measures the inference speed.

Method	Infer. Memory (GB) ↓	Speed (FPS) ↑
CleanPose [52]	10.04	24.33
Diff9D [65]	9.60	17.20
GenCOPE (ours)	**2.89**	**40.68**

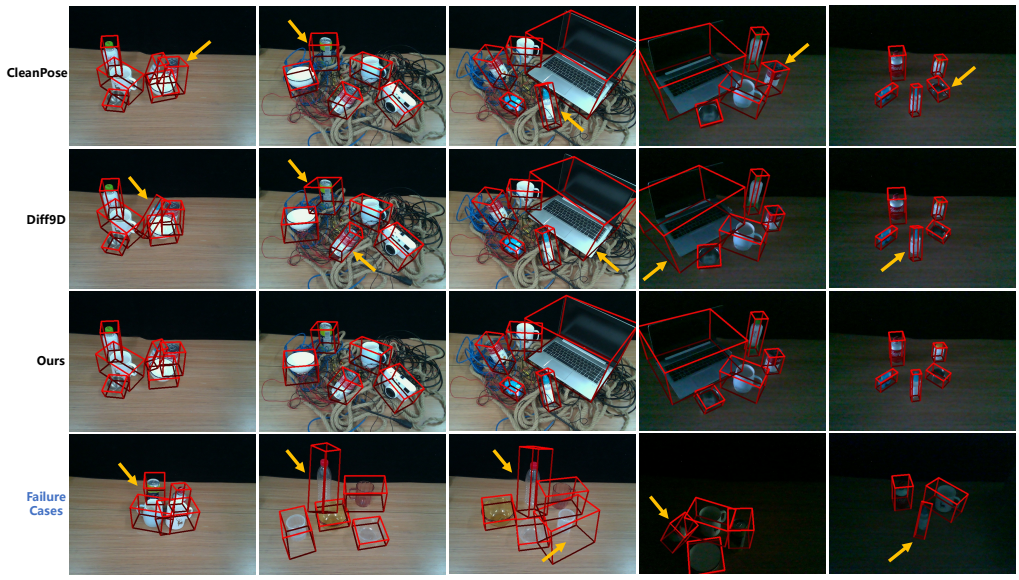

Figure 4: Qualitative comparison with CleanPose [52] and Diff9D [65] in real-world robotic scenarios.

inference speed. This efficiency stems from the use of lightweight 2D and 3D encoders, the exclusive reliance on global features during 2D-3D cross consistency learning, and a compact pose-decoupled decoder. As a result, GenCOPE exhibits strong deployability and high efficiency, making it well-suited for real-world robotic platform, as mentioned in Implementation Details.

Qualitative Analysis We deploy the trained, lightweight GenCOPE model on our robotic platform (with only 6 GB GPU memory) for on-board inference. In contrast, due to the higher GPU memory requirements of CleanPose [52] and Diff9D [65], the captured RGB-D images are transmitted to the same workstation used during training for inference. For a fair comparison, all methods rely on identical object segmentation results obtained from Mask R-CNN [66].

Qualitative Analysis on REAL275 and Wild6D Datasets: Figure 3 presents qualitative comparisons between GenCOPE with CleanPose [52] and Diff9D [65] on the REAL275 dataset [2] and Wild6D dataset [59], respectively. On the REAL275 dataset (first two rows), compared with CleanPose, GenCOPE produces more accurate 3D bounding boxes under cluttered scenes. In particular, CleanPose suffers from noticeable rotation and size deviations, as highlighted by the yellow arrows, whereas GenCOPE yields tighter alignment with ground-truth poses. On the Wild6D dataset (last two rows), which features diverse real-world scenes with complex backgrounds and severe domain shifts, GenCOPE consistently outperforms Diff9D. While Diff9D occasionally exhibits misaligned orientations, GenCOPE maintains stable pose predictions across different object categories. These results show that explicitly enforcing semantic consistency across modalities enables GenCOPE to suppress domain-specific cues better and achieve stronger domain generalization.

Qualitative Analysis in Real World: Figure 4 further evaluates GenCOPE in real-world robotic scenarios, where multiple objects are placed in occluded, cluttered, and low-light environments. Compared with CleanPose [52] and Diff9D [65], GenCOPE produces more accurate pose estimations. These results show that the compared methods tend to produce incorrect orientations in complex scenes, while GenCOPE delivers more reliable predictions. In addition, the last row of Fig. 4 illustrates some representative failure cases, which mainly arise under severe occlusion, transparent objects with incomplete depth observations, or extremely low-light conditions (these scenarios still have significant domain gaps compared to the training data). Despite some limitations, GenCOPE demonstrates improved robustness compared to existing methods. Overall, these real-world results validate the practical effectiveness and deployability of GenCOPE for robotic tasks, aligning well with the quantitative evaluations reported earlier. **Some robotic picking demos** are provided in the Appendix B and the supplementary video.

4.2 Ablation Studies

We conduct ablations on two key components to quantify their contributions to GenCOPE on REAL275 dataset [2] (Table 4): (i) the 2D/3D semantic consistency encoders, the loss of semantic

consistency supervision, and the choice of LLM (Rows A); and (ii) the 2D-3D cross consistency learning module and the self-/cross-attention mechanisms within the 2D-3D cross transformer (Rows B). **More ablation studies** on the use of text feature F_{text} with or without 2D/3D semantic consistency supervision are provided in the Appendix D.

2D/3D Semantic Consistency Encoders First, we remove semantic consistency supervision from either encoder, which consistently degrades performance (Table 4 A0 and A1), confirming its necessity for robust representation learning. In particular, removing 2D consistency causes a larger regression (e.g., $38.1\% \rightarrow 34.0\%$ at $5°2\text{cm}$) than removing the 3D counterpart ($38.1\% \rightarrow 36.5\%$), suggesting that 2D features are more vulnerable to the Syn2Real domain gap and thus benefit more from explicit semantic alignment. Further, we conduct a finer-grained ablation by replacing part-structured descriptions with class-name-only text (e.g., "camera", "laptop", etc) for both 2D and 3D encoders, and observe a clear performance drop (Table 4 A2). This result confirms that the improvement arises from the structured semantic design rather than class-level supervision, providing direct evidence that intra-class semantic consistency alleviates the Syn2Real domain gap. Next, we replace our loss with a cosine similarity loss $\mathcal{L}_{cos}$ (Table 4 A3), which further incurs a substantial drop in all metrics, indicating that our formulation provides more effective supervision within a training batch containing multiple positive samples. Finally, we replace Gemini-3 [36] with Qwen2.5-VL-7B [74] and GPT-4 [75] while keeping the semantic query template and all other settings unchanged, generating comparable accuracy (Table 4 A4 and A5 vs. C0), which highlights that GenCOPE is robust to the choice of LLM.

Table 4: Ablation studies for the 2D/3D semantic consistency encoders and 2D-3D cross consistency learning module on the REAL275 dataset [2].

Row	Module Removal	mean Average Precision (mAP)				
		$3D_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
A0	2D Semantic Consistency	42.6	34.0	46.2	52.8	72.1
A1	3D Semantic Consistency	43.1	36.5	47.2	54.6	73.3
A2	Class-Name-Only Text	41.9	33.4	45.8	52.1	71.7
A3	$\mathcal{L}_i \rightarrow \mathcal{L}_{cos}$	41.7	33.1	45.0	51.6	70.9
A4	Gemini-3 [36] → Qwen2.5-VL-7B [74]	43.9	37.7	49.4	56.8	74.3
A5	Gemini-3 [36] → GPT-4 [75]	44.3	38.0	49.6	**56.9**	**74.6**
B0	2D-3D Cross Consistency	42.9	34.9	47.0	53.8	72.9
B1	SelfAttn(·) in (e) of Fig. 2	44.1	37.3	48.5	55.6	73.6
B2	CrossAttn(·) in (e) of Fig. 2	43.3	36.0	47.3	54.7	73.1
C0	Original GenCOPE	**44.5**	**38.1**	**49.9**	56.4	74.4

2D-3D Cross Consistency Learning When the module is completely removed and the 2D and 3D features are directly concatenated for pose regression (Table 4 B0), the accuracy drops across all metrics, validating the importance of cross-modal fusion. We then separately ablate the self-/cross-attention components within the 2D-3D cross transformer (Table 4 B1 and B2), both of which result in degradation. These results validate the necessity of the proposed module design and demonstrate that explicitly modeling cross-modal interactions between 2D and 3D features is crucial for accurate pose estimation.

5 Conclusion

We presented GenCOPE, a simple yet effective Syn2Real generalized COPE method that enables models trained solely on synthetic data to generalize robustly to real-world deployment. We first analyze the root causes of domain discrepancies between synthetic and real-world data, and guided by these findings, introduced semantic consistency-guided 2D and 3D encoders and the 2D-3D cross consistency learning network with an end-to-end pose-decoupled decoder. Extensive results show that GenCOPE achieves superior domain generalization with minimal computational overhead and real-time inference, making it well-suited for practical integration into real-world robotic platforms.

Limitation Although GenCOPE achieves encouraging results, the current formulation does not explicitly account for pose equivalence in strongly symmetric categories, where multiple rotations can describe the same physical configuration. Incorporating symmetry-aware losses and evaluation protocols remains an important direction for future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants 62473141 and U22A2059, and also supported by the National Science and Technology Major Project of China under Grant 2026ZD1610900.

References

[1] Jian Liu, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. Deep learning-based object pose estimation: a comprehensive survey. *International Journal of Computer Vision*, 134(2):81, 2026. 2

[2] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2, 3, 6, 7, 9, 10, 18

[3] Ruiqi Wang, Xinggang Wang, Te Li, Rong Yang, Minhong Wan, and Wenyu Liu. Query6dof: Learning sparse queries as implicit shape prior for category-level 6dof pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14055–14064, 2023. 2

[4] Linfang Zheng, Tze Ho Elden Tse, Chen Wang, Yinghan Sun, Hua Chen, Ales Leonardis, Wei Zhang, and Hyung Jin Chang. Georef: Geometric alignment across shape variation for category-level object pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10693–10703, 2024. 2

[5] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21231, 2022. 2

[6] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22498–22508, 2024. 2

[7] Hui Yang, Wei Sun, Jian Liu, Jin Zheng, Zhenqi Dai, and Ajmal Mian. Lancope: Language-guided category-level object pose estimation from a single rgb image. *IEEE Robotics and Automation Letters*, 2025. 2

[8] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216, 2024. 2

[9] Ziqin Huang, Gu Wang, Chenyangguang Zhang, Ruida Zhang, Xiu Li, and Xiangyang Ji. Givepose: Gradual intra-class variation elimination for rgb-based category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22055–22066, 2025. 2

[10] Jian Liu, Wei Sun, Kai Zeng, Jin Zheng, Hui Yang, Hossein Rahmani, Ajmal Mian, and Lin Wang. Scalable unseen objects 6-dof absolute pose estimation with robotic integration. *IEEE Transactions on Robotics*, 42:1884–1901, 2026. 2

[11] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3343–3352, 2019. 2

[12] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11641, 2020. 2

[13] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3003–3013, 2021. 2

[14] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 2

[15] Hui Yang, Wei Sun, Jian Liu, Jin Zheng, Jian Xiao, and Ajmal Mian. Occlusion-aware 3d hand-object pose estimation with masked autoencoders. *IEEE Transactions on Multimedia*, 2026. 2

[16] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *European Conference on Computer Vision*, pages 139–156, 2020. 2

[17] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4807–4814, 2021. 2

[18] Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In *European Conference on Computer Vision*, pages 220–236, 2022. 2

[19] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 8067–8074, 2021. 2

[20] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. Catre: Iterative point clouds alignment for category-level object pose refinement. In *European Conference on Computer Vision*, pages 499–516, 2022. 2, 6, 7

[21] Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21040–21049, 2024. 2, 3

[22] Jian Liu, Wei Sun, Hui Yang, Jin Zheng, Zichen Geng, Hossein Rahmani, and Ajmal Mian. Monodiff9d: Monocular category-level 9d object pose estimation via diffusion model. In *IEEE International Conference on Robotics and Automation*, pages 8687–8694, 2025. 2

[23] Yifan Yang, Peili Song, Enfan Lan, Dong Liu, and Jingtai Liu. Mk-pose: Category-level object pose estimation via multimodal-based keypoint learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4232–4239, 2025. 2

[24] Hui Yang, Wei Sun, Jian Liu, Jian Xiao, Tao Xie, Hossein Rahmani, Ajmal Mian, Nicu Sebe, and Gim Hee Lee. Genhoi: Generalized hand-object pose estimation with occlusion awareness. In *European Conference on Computer Vision*, pages 684–701, 2026. 2

[25] Huan Ren, Yihan Chen, Chuxin Wang, Nailong Liu, Wenfei Yang, and Tianzhu Zhang. Compose: A unified completion-pose framework for robust category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 2

[26] Xiao Lin, Minghao Zhu, Yun Peng, Liuyi Wang, Qiyi Wang, Chengju Liu, and Qijun Chen. Mempose: Category-level object pose estimation with memory. In *European Conference on Computer Vision*, pages 156–174, 2026. 2

[27] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11076–11084, 2020. 2

[28] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2021. 2

[29] Hui Yang, Wei Sun, Jian Liu, Jin Zheng, Zhiwen Zeng, and Ajmal Mian. Rgb-based category-level object pose estimation via depth recovery and adaptive refinement. *IEEE Robotics and Automation Letters*, 2025. 2

[30] Zichen Geng, Zeeshan Hayder, Bo Miao, Jian Liu, Wei Liu, and Ajmal Mian. Disentangled hierarchical vae for 3d human-human interaction generation. In *International Conference on Learning Representations*, 2026. 2

[31] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 5, 18

[33] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *European Conference on Computer Vision*, pages 19–34, 2022. 2, 3, 6, 7

[34] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13978–13988, 2023. 2, 3, 6

[35] Bo Miao, Weijia Liu, Jun Luo, Lachlan Shinnick, Jian Liu, Thomas Hamilton-Smith, Yuhe Yang, Zijie Wu, Vanja Videnovic, Feras Dayoub, and Anton van den Hengel. Langmap: A human-verified benchmark for hierarchical open-vocabulary goal navigation. In *Advances in Neural Information Processing Systems*, 2026. 2

[36] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 4, 10, 16

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 4, 18

[38] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022. 2

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 4, 17

[40] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2, 4, 17

[41] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pages 530–546, 2020. 3, 7, 8

[42] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021. 3, 6, 7, 8

[43] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *International Conference on Robotics and Automation*, pages 10632–10640, 2022. 3

[44] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pages 655–672, 2022. 3

[45] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7452–7459, 2022. 3

[46] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 3

[47] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021. 3

[48] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 3, 8

[49] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 1–14, 2023. 3, 6

[50] Xiao Lin, Minghao Zhu, Ronghao Dang, Guangliang Zhou, Shaolong Shu, Feng Lin, Chengju Liu, and Qijun Chen. Clipose: Category-level object pose estimation with pre-trained vision-language knowledge. *arXiv preprint arXiv:2402.15726*, 2024. 3

[51] Huan Ren, Wenfei Yang, Shifeng Zhang, and Tianzhu Zhang. Rethinking correspondence-based category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1170–1179, 2025. 3

[52] Xiao Lin, Yun Peng, Liuyi Wang, Xianyou Zhong, Minghao Zhu, Yi Feng, Jingwei Yang, Chengju Liu, and Qijun Chen. Cleanpose: Category-level object pose estimation via causal learning and knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5990–6000, 2025. 3, 6, 7, 8, 9

[53] Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14001–14011, 2023. 3, 7

[54] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 3, 6, 7

[55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[56] Weihang Li, Hongli Xu, Junwen Huang, Hyunjun Jung, Peter KT Yu, Nassir Navab, and Benjamin Busam. Gce-pose: Global context enhancement for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27154–27165, 2025. 3

[57] Yang You, Ruoxi Shi, Weiming Wang, and Cewu Lu. Cppf: Towards robust category-level 9d pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6866–6875, 2022. 3

[58] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2022. 3, 7

[59] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 27469–27483, 2022. 3, 6, 7, 8, 9

[60] Wanli Peng, Jianhang Yan, Hongtao Wen, and Yi Sun. Self-supervised category-level 6d object pose estimation with deep implicit shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2082–2090, 2022. 3, 7

[61] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. Uda-cope: unsupervised domain adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14891–14900, 2022. 3, 7

[62] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. Tta-cope: Test-time adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21285–21295, 2023. 3, 7

[63] Jian Liu, Wei Sun, Chongpei Liu, Hui Yang, Xing Zhang, and Ajmal Mian. Mh6d: Multi-hypothesis consistency learning for category-level 6-d object pose estimation. *IEEE Transactions on Neural Networks and Learning Systems.*, 36(3):4820–4833, 2025. 3

[64] Takuya Ikeda, Sergey Zakharov, Tianyi Ko, Muhammad Zubair Irshad, Robert Lee, Katherine Liu, Rares Ambrus, and Koichi Nishiwaki. Diffusionnocs: Managing symmetry and uncertainty in sim2real multi-modal category-level pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7406–7413, 2024. 3, 7

[65] Jian Liu, Wei Sun, Hui Yang, Pengchao Deng, Chongpei Liu, Nicu Sebe, Hossein Rahmani, and Ajmal Mian. Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5520–5537, 2025. 3, 6, 7, 8, 9

[66] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 3, 6, 7, 9

[67] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 3

[68] Bo Miao, Mingtao Feng, Zijie Wu, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Referring human pose and mask estimation in the wild. *Advances in Neural Information Processing Systems*, 37:44791–44813, 2024. 3

[69] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, 2020. 5, 6

[70] Sida Peng, Xiaowei Zhou, Yuan Liu, Haotong Lin, Qixing Huang, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3212–3223, 2022. 5

[71] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision*, pages 464–472, 2017. 6

[72] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, and Qiang Fu. Robotic continuous grasping system by shape transformer-guided multi-object category-level 6-d pose estimation. *IEEE Transactions on Industrial Informatics*, 19(11):11171–11181, 2023. 7

[73] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[74] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 10

[75] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 10

Appendix

A Additional Semantic Samples Generated by LLM

As illustrated in Fig. A, the LLM [36] generates diverse semantic descriptions for different object categories. Even when queried with an identical template for the same category, the LLM produces varied responses across multiple calls. To exploit this variability, we query the LLM ten times per object category during training and randomly sample one description at each iteration, thereby increasing the diversity and stochasticity of semantic supervision.

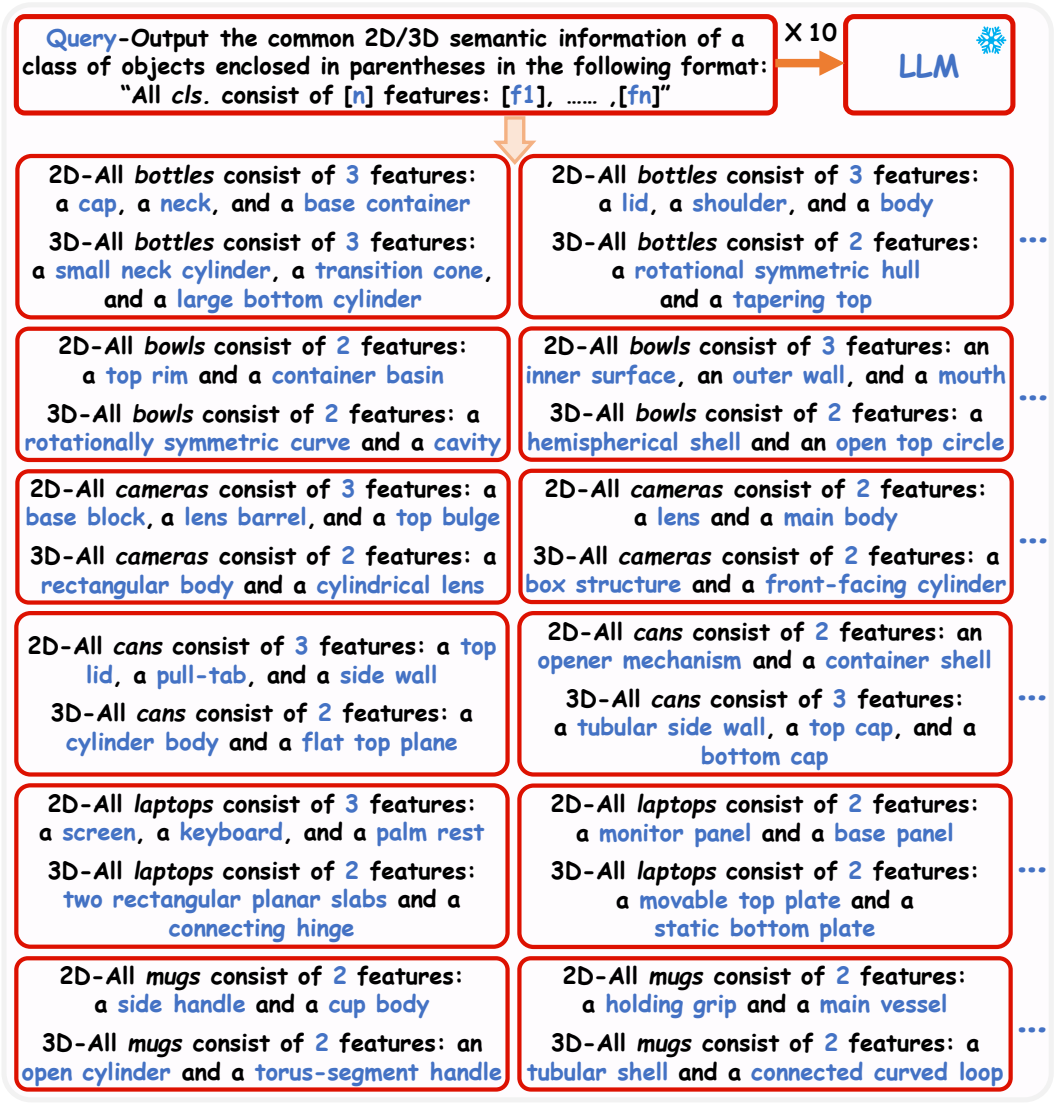

Figure A: Diverse semantic samples generated by LLM for different object categories.

B GenCOPE Guided Robotic Picking

To validate the practicality of GenCOPE trained solely on synthetic data, we deploy it on a real-world eye-in-hand robotic picking platform equipped with an Intel RealSense L515 RGB-D camera and an onboard NVIDIA GTX 1660 GPU with 6 GB memory. Owing to its lightweight design, GenCOPE runs entirely on the embedded hardware without offloading computation. Representative pose estimation results are shown in the top-left of Fig. B. To ensure safe grasp execution, the

grasp position is manually offset by $2cm$ above the tabletop, and the grasp orientation is tilted upward by 30°. Multiple objects in the scene, including textureless metallic components, are then sequentially picked based on their estimated center depths, as illustrated in Fig. B. These real-world experiments demonstrate the strong deployability, high efficiency, and robust Syn2Real generalization of GenCOPE, confirming its effectiveness for robotic picking.

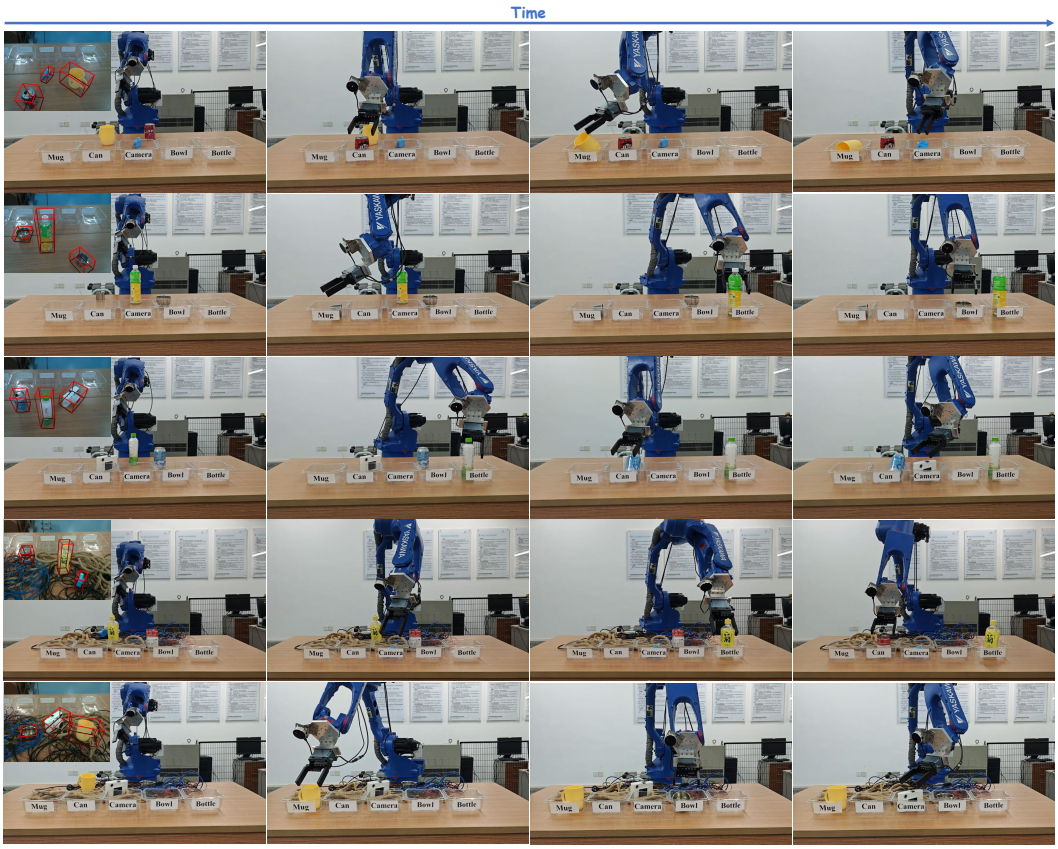

Figure B: Visualization of GenCOPE-guided multi-object continuous robotic picking.

C Details of Some Modules

In the feature encoding stage, we adopt ResNet18 [39] with the final fully connected layer removed to extract 2D features from the input RGB image. For the point cloud back-projected from the depth map, we employ a PointNet-based encoder (detailed in Fig. C.1); unlike the standard PointNet [40], it is used exclusively for global feature extraction to prioritize deployability and computational efficiency. In addition, both the self-attention and cross-attention modules in the 2D-3D cross consistency learning network are implemented using a shared transformer block, as illustrated in Fig. C.2.

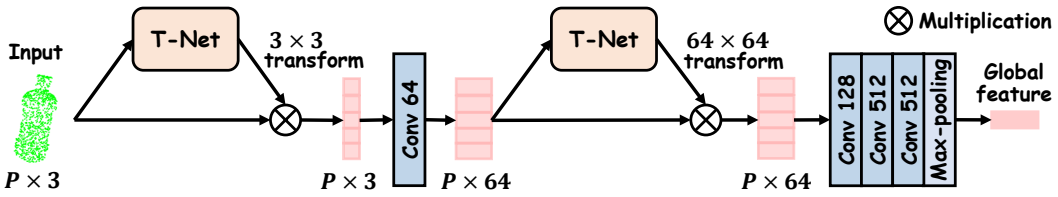

Figure C.1: Detailed structure of the PointNet-based [40] point cloud global feature encoder.

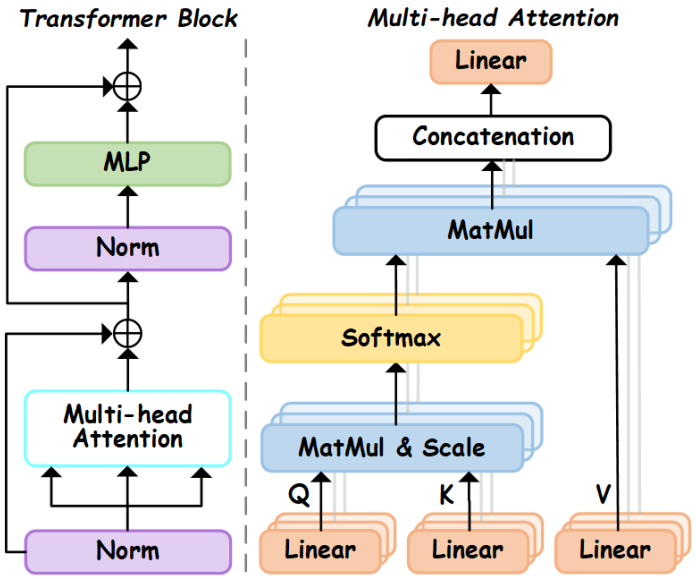

Figure C.2: Detailed structure of the transformer block [32] used in self-attention and cross-attention modules. $\oplus$ and "MatMul" represent residual connection and matrix multiplication, respectively.

D Additional Ablation Studies

Table A: Ablation studies on the use of text features F_{text} (including $F_{\text{text-2D}}$ and $F_{\text{text-3D}}$), with and without 2D/3D semantic consistency supervision, on the REAL275 dataset [2].

Row	F_{text} in Train. & Infer.	$\mathcal{L}_{C_text\text{-}2D}$ & $\mathcal{L}_{C_text\text{-}3D}$	mean Average Precision (mAP)				
			$3D_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
S0	residual	–	42.1	33.7	45.4	51.6	71.0
S1	concatenate	–	42.5	34.6	46.1	52.5	72.1
S2	residual	✓	44.0	38.0	49.6	55.9	74.0
S3	concatenate	✓	44.4	**38.3**	**50.0**	56.2	**74.8**
Ours	–	✓	**44.5**	38.1	49.9	**56.4**	74.4

Table A presents an ablation study analyzing the role of text features and the proposed 2D/3D semantic consistency supervision on the REAL275 dataset [2]. The second column evaluates whether CLIP-based [37] text features are introduced during both training and inference via direct residual addition or feature concatenation, while the third column examines the effect of enabling the proposed 2D and 3D semantic consistency supervision during training. Specifically, the direct residual addition is as follows:

$$\begin{aligned} F_{C\text{-}2D} &= F_{C\text{-}2D} + F_{text\text{-}2D}, \\ F_{C\text{-}3D} &= F_{C\text{-}3D} + F_{text\text{-}3D}. \end{aligned} \tag{8}$$

Then, $F_{C\text{-}2D}$ and $F_{C\text{-}3D}$ are fed into the same 2D-3D cross consistency learning module for pose regression. The feature concatenation is defined as follows:

$$\begin{aligned} F_{C\text{-}2D} &= \text{Concat}(F_{C\text{-}2D},\ F_{text\text{-}2D}), \\ F_{C\text{-}3D} &= \text{Concat}(F_{C\text{-}3D},\ F_{text\text{-}3D}). \end{aligned} \tag{9}$$

Based on this, we double the feature dimensions of the 2D-3D cross consistency learning module and the pose decoder, while preserving the same network architecture, to align with the corresponding $F_{C\text{-}2D}$ and $F_{C\text{-}3D}$. The experimental results in Table A show that directly injecting text features into the visual encoders does not consistently improve performance (Table A S0 and S1). Feature concatenation achieves performance comparable to our full model (Table A S3), whereas residual

addition slightly degrades (Table A S2). Moreover, introducing text features during both training and inference inevitably increases computational cost and memory consumption, which is undesirable for real-time robotic deployment. In contrast, incorporating the proposed semantic consistency supervision consistently improves performance, regardless of whether text features are explicitly fused into the 2D/3D representations. These results indicate that enforcing semantic alignment during feature learning is more effective than directly multimodal feature fusion. Overall, the findings validate that the proposed 2D/3D semantic consistency supervision enables the encoders to focus on domain-invariant semantic cues, thereby improving robustness to domain-specific information while maintaining computational efficiency and training stability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the main contributions of GenCOPE, which are supported by the methods and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Figure 4 explicitly discusses the failure cases and Section 5 discusses the limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Details can be found in the Method, Experiments, and Appendix sections.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details can be found in the Method, Experiments, and Appendix sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code will be released after the paper is accepted.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details can be found in the Experiments and Appendix sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Details can be found in the Experiments and Appendix sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details can be found in the Experiments sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics, as it does not involve human subjects, sensitive data, or high-risk model/data release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work does not have societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work does not release high-risk models or datasets; no special safeguards are required for responsible use.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and pre-trained models used are properly credited to the original creators, and their usage respects the intended terms as described in the original papers.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release all configuration files used in the experiments, along with documentation on how to use them to reproduce the results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve research with human subjects or require IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use Gemini-3 to generate category-level 2D and 3D semantic descriptions, which are encoded by CLIP and used as semantic consistency supervision during training. We additionally evaluate alternative language models, including Qwen2.5-VL-7B and GPT-4, in the ablation study.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.