GROUND2POSE: GEOMETRIC SELF-VERIFICATION FOR TRAINING-FREE OPEN-VOCABULARY 6D POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the 6D pose of novel objects is critical for open-world embodied perception, yet existing estimators rely heavily on object CAD models, multi-view calibration, or task-specific fine-tuning. While emerging language-guided methods bypass pre-scanned 3D assets, they tightly couple 2D localization with 3D pose recovery, allowing upstream segmentation errors to catastrophically corrupt downstream geometric reasoning. We present Ground2Pose, a training-free framework that breaks these constraints by estimating 6D relative object pose from an anchor–query RGB-D view pair and a text description, without CAD assets, reference calibration, or task supervision. Ground2Pose explicitly decouples open-vocabulary 2D localization from 3D geometric reasoning. At its core is the Geometric Self-Verification Solver (GVS)—a solver that turns its own closed-form alignment into an outlier verifier, systematically pruning correspondence outliers via 3D alignment residuals rather than unreliable semantic or heuristic confidence scores. On four diverse benchmarks under our strict single-anchor, predicted-mask protocol, Ground2Pose achieves SOTA performance (70.9 ADD(S)-0.1d and 42.2 BOP recall), substantially outperforming the strongest comparable competitor by +40.8 and +8.8 points. Ablations and controlled mask interventions reveal that while upstream localization sets the empirical accuracy ceiling, correspondence-level geometric verification determines how closely the solver approaches it, establishing a robust *verification-over-reweighting* design principle. With fixed pre-trained backbones and closed-form efficiency, Ground2Pose provides a reproducible and robust foundation for open-world spatial grounding. Code will be made publicly available upon acceptance.

1 INTRODUCTION

Estimating the 6D pose of objects from RGB-D observations is fundamental to robotic manipulation, augmented reality, and autonomous inspection Xiang et al. (2018); Peng et al. (2019); Zeng et al. (2017); Du et al. (2021); He et al. (2021); Wang et al. (2019a); Cai et al. (2022); Liu et al. (2026). While supervised methods achieve high accuracy for known objects, they rely on object-specific training data, pose annotations, or CAD models, limiting their applicability to novel objects. Language-guided open-vocabulary methods Corsetti et al. (2024; 2026); Tu et al. (2025); Deng et al. (2021) offer a transformative paradigm: arbitrary target objects can be specified via free-form text descriptions, enabling open-world embodied perception without offline object modeling. However, this formulation introduces a fundamental dependency: **the geometric backend can only reason over the object evidence delivered by the localization stage**, and imperfect masks corrupt downstream correspondences and limit attainable pose accuracy.

We study the localization–geometry interface in a controlled setting. Across multiple off-the-shelf open-vocabulary segmenters, localization quality strongly predicts downstream pose accuracy, with mask IoU and ADD(S)-0.1d exhibiting a rank correlation of $\rho = 0.934$ (Table 3). Changing only the mask source can alter downstream ADD(S)-0.1d by up to $\sim$36 points. These observations are corroborated by controlled mask interventions: replacing predicted masks with ground truth (Table 3, oracle row) and varying the solver under fixed mask quality (Appendix Table 16) turn the correlation into intervention-level evidence. The resulting design principle: **localization determines**

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

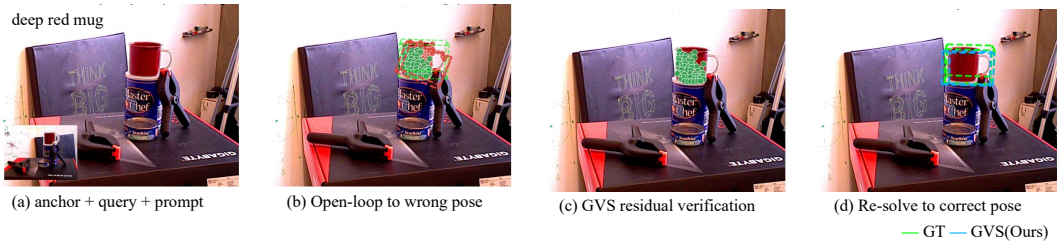


Figure 1: **Motivation: even with a correct mask, open-loop recovery fails on correspondence-level outliers.** A real YCB-Video example. (a) Anchor view, query scene and prompt "deep red mug". (b) Among the 87 mask-filtered correspondences, 26 are geometric outliers (red), three of them extreme cross-instance matches ($>$1.25 m) projected onto a neighboring object; the open-loop solve over all matches is dragged off target (red dashed box, ADD 41.4 mm). (c) GVS ranks matches by their 3D alignment residuals and trims the worst fraction (γ=0.7, 87$\rightarrow$61). (d) A single re-solve on the retained set realigns the prediction (blue dashed) with the ground truth (green dashed), reducing ADD to 15.3 mm (2.7$\times$).

the empirical accuracy ceiling, while correspondence-level geometric verification determines how closely the solver approaches it.

Motivated by this insight, we present Ground2Pose, a training-free open-vocabulary 6D pose estimation framework that explicitly separates *language-guided localization* from *geometric pose recovery*. Given a text description and an anchor–query RGB-D frame pair, Ground2Pose first localizes the target object using a modular open-vocabulary segmentation module (default: Grounded-SAM Ren et al. (2024)), then performs mask-constrained dense matching, closed-form pose recovery, and geometric self-verification. At the core of the geometric stage is the *geometric self-verification solver* (GVS), which identifies unreliable correspondences through their 3D alignment residuals (Fig. 1) and re-solves the pose in closed form. The framework requires no CAD model, no multi-view reference set, and no task-specific training.

In summary, our main contributions are:

- We provide a comprehensive analysis of 6D pose estimation paradigms across CAD-based, reference-view, and language-guided regimes, and propose a decoupled, training-free framework that separates language-guided 2D localization from 3D metric pose recovery, breaking the fragile dependency of monolithic pipelines and empowering the geometric backend to independently verify correspondences and generalize to novel objects without CAD assets or pose supervision.
- We provide a deterministic solver for correspondence-level pose recovery, termed the Geometric Self-Verification Solver (GVS), which turns its closed-form alignment into an outlier verifier, pruning outliers directly through 3D alignment residuals rather than fragile semantic confidences, resulting in consistent improvements of up to +4.25 ADD(S)-0.1d points under default localization and up to +17.35 points under noisy masks compared to open-loop estimation.
- We adopt a unified anchor–query evaluation protocol across four standardized benchmarks (YCB-Video, NOCS-REAL275, Toyota Light, and LineMOD) on 2,000 pairs each, demonstrating state-of-the-art performance (+40.8 ADD(S)-0.1d over the strongest practical competitor) and formulate a robust "verification-over-reweighting" design principle to guide future training-free spatial grounding systems.

2 Related Work

Table 1 positions contemporary 6D pose estimation frameworks along two axes: the reference asset each method requires at inference (*Reference Asset*), and the supervision regime it assumes (whether it is *training-free* and/or *open-vocabulary*). We focus particularly on CAD-free and language-driven paradigms, which share our practical objective of novel-object manipulation without pre-scanned 3D assets or task-specific fine-tuning.

Table 1: **Assumption matrix of compared methods.** ✓ indicates that the method satisfies the property without heavy priors or supervision; *Reference Asset* denotes the reference data each method needs at inference. Ground2Pose is the only framework that simultaneously achieves **CAD-free**, **training-free**, and **open-vocabulary** 6D pose estimation using only a single anchor RGB-D frame.

Method	Modality	Reference Asset	CAD-Free	Training-Free	Open-Vocab
Ground2Pose (Ours)	RGB-D	Single Anchor	✓	✓	✓
Oryon (2024)	RGB-D	Single Anchor	✓	✗	✓
Horyon (2026)	RGB-D	Single Anchor	✓	✗	✓
LatentFusion (2020)	RGB-D	Reference Views	✓	✗	✗
OnePoseviaGen (2025)	RGB	Anchor + Gen. Views	✓	✗	✗
GigaPose (2024)	RGB	CAD Templates	✗	✓	✗
FoundationPose (2024)	RGB-D	CAD Model	✗	✗	✗
SAM-6D (2024b)	RGB-D	CAD Model	✗	✗	✓
FreeZe (2024)	RGB-D	CAD Model	✗	✓	✗
Geo6DPose (2025)	RGB-D	CAD Templates	✗	✓	✗
PoseDiffusion (2023)	RGB	Sparse Views	✓	✗	✗
RelPose++ (2024a)	RGB	Sparse Views	✓	✗	✗

Prior work spans several regimes (Table 1). We discuss the most relevant ones in turn, focusing on the assumptions that limit open-world deployment.

Training-free 6D pose estimation. CAD-reliant methods Drost et al. (2010); Caraffa et al. (2024); Villena Toro & Tarkian (2025) match against pre-existing 3D meshes or rendered templates, achieving geometric accuracy but requiring explicit CAD models unavailable for novel objects. Reference-based CAD-free methods Liu et al. (2022); Sun et al. (2022); Deng et al. (2025); Lee et al. (2025) bypass 3D geometry via visual correspondences but require dense multi-view sequences or pre-built SfM maps, and lack language understanding. The combination of single-anchor, CAD-free, training-free, and language-guided pose estimation remains largely unaddressed.

Language-guided pose estimation. Oryon Corsetti et al. (2024) and Horyon Corsetti et al. (2026) pioneered language-conditioned localization bridged with geometric registration. However, they suffer two key limitations: (1) reliance on task-specific pose supervision or fine-tuned segmentation heads, hindering zero-shot generalization; and (2) open-loop pipelines where upstream mask errors irreversibly propagate without geometric verification. Ground2Pose rethinks this via a fully training-free, decoupled architecture—treating language grounding as a plug-and-play front-end and introducing GVS as a deterministic backend that actively prunes localization-induced outliers, establishing a *verification-over-reweighting* design principle.

Robust registration. RANSAC Fischler & Bolles (1981) exhibits stochastic variance and requires large hypothesis budgets under dense correspondences. Ground2Pose's GVS inherits deterministic trimmed least-squares efficiency Chetverikov et al. (2002), using the full-set closed-form SVD as a globally stable verifier with negligible compute cost.

Summary. Prior works either rely on CAD models, require task-specific training, or lack explicit geometric verification against upstream semantic errors. Ground2Pose addresses this gap by unifying modular open-vocabulary grounding with deterministic correspondence self-verification into a coherent, training-free framework.

3 Framework and Design Rationale

We define key terms: *Training-free* means no component receives task-specific (pose) supervision—all modules use publicly released weights with only fixed hyperparameters tuned. *Novel objects* are unseen at test time: no CAD model, no reference views, no object-specific training. *Open-vocabulary* localization means the target is specified only by a free-form text phrase at inference time.

Given an anchor image I_a, a query image I_q, and a text description T specifying the target object, we estimate the rigid transformation $[\mathbf{R}|\mathbf{t}]$ from I_a to I_q. Our setting assumes a single anchor RGB-D frame with its pose, plus a text phrase; it requires no CAD model, no multi-view reference,

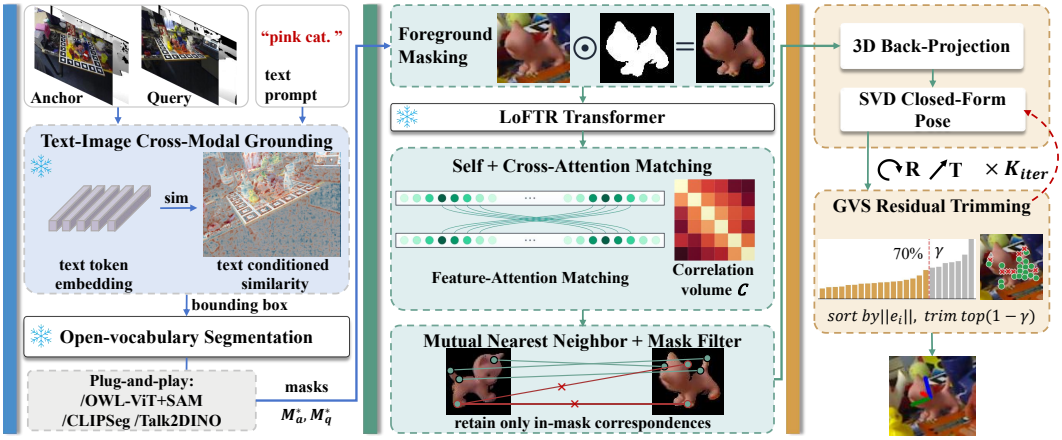


Figure 2: **Framework overview of Ground2Pose.** Given an anchor RGB-D image, a query RGB-D image, and a text description, the language-guided localization stage predicts the target mask in each image; this stage is a modular interface (Sec. 3.1) instantiated by off-the-shelf open-vocabulary segmenters. Mask-constrained dense matching (Sec. 3.2) establishes correspondences, and 3D lifting with SVD, refined by the GVS verification loop (Sec. 3.3), recovers the final pose. The learned components remain frozen and are used off the shelf, enabling open-vocabulary 6D pose estimation without pose-specific training or fine-tuning.

and no task-specific training. As illustrated in Fig. 2, our framework consists of three stages: (1) language-guided object localization, (2) mask-constrained dense matching, and (3) geometry-aware pose recovery with the geometric self-verification solver (GVS). The localization stage is modular: any open-vocabulary segmenter mapping (I, T) to a binary mask can be plugged in; our default uses Grounded-SAM Ren et al. (2024).

3.1 Language-Guided Localization

The first stage grounds the target object from natural-language instructions alone. We treat it as a modular interface: any open-vocabulary segmenter mapping (I, T) to a binary mask can be plugged in, applied independently to I_a and I_q to produce masks M_a^* and M_q^*. No downstream component depends on how these masks were obtained, allowing us to isolate the influence of localization quality on final pose accuracy (Sec. 4). Our default instantiates this interface with off-the-shelf Grounded-SAM Ren et al. (2024): Grounding-DINO Liu et al. (2024) detects the best-matching bounding box, which prompts SAM Kirillov et al. (2023) to produce a pixel-accurate mask. We also evaluate CLIPSeg Lüddecke & Ecker (2022), OWL-ViT+SAM Minderer et al. (2022); Kirillov et al. (2023), and Talk2DINO Barsellotti et al. (2025) as alternative instantiations.

3.2 Mask-Constrained Dense Matching

With the localization masks M_a^* and M_q^*, we perform dense correspondence estimation. We first apply the masks to suppress background before feature extraction:

$$\tilde{I}_a = I_a^{\text{gray}} \odot M_a^*, \quad \tilde{I}_q = I_q^{\text{gray}} \odot M_q^*, \tag{1}$$

where $\odot$ denotes element-wise multiplication. We then use LoFTR Sun et al. (2021) as an off-the-shelf dense matcher on the masked grayscale images to obtain correspondences with confidence scores $\{(\mathbf{p}_a^i, \mathbf{p}_q^i, c^i)\}_{i=1}^N$. Correspondences are retained only if both endpoints fall inside the predicted masks:

$$\mathcal{C}_{\text{geo}} = \left\{(\mathbf{p}_a^i, \mathbf{p}_q^i, c^i) \;\middle|\; M_a^*(\mathbf{p}_a^i) = 1 \;\wedge\; M_q^*(\mathbf{p}_q^i) = 1\right\}. \tag{2}$$

Optionally, correspondences can be ranked by confidence score to retain the K_{top} highest matches ($K_{\text{top}} = $ all by default; Appendix B).

3.3 GVS: Geometry-Aware Pose Recovery with Self-Verification

3D Lifting and Rigid Alignment. The filtered 2D correspondences are back-projected to 3D using depth maps D_a, D_q and camera intrinsic matrices $\mathbf{K}_a, \mathbf{K}_q \in \mathbb{R}^{3\times3}$:

$$\mathbf{P}_{ai} = z_{ai}\mathbf{K}_a^{-1}[u_{ai}, v_{ai}, 1]^\top, \quad \mathbf{P}_{qi} = z_{qi}\mathbf{K}_q^{-1}[u_{qi}, v_{qi}, 1]^\top, \tag{3}$$

where $\mathbf{p}_a^i = (u_{ai}, v_{ai})^\top$, $\mathbf{p}_q^i = (u_{qi}, v_{qi})^\top$, and the depths $z_{ai} = D_a(\mathbf{p}_a^i)$, $z_{qi} = D_q(\mathbf{p}_q^i)$ are sampled from the respective depth maps. The optimal rigid transformation $(\mathbf{R}, \mathbf{t})$ is recovered via SVD-based closed-form solution Kabsch (1976); Umeyama (1991). Let $M = |\mathcal{C}_{\text{geo}}|$ denote the number of lifted correspondences entering the solver, let $\bar{\mathbf{P}}_a = \frac{1}{M}\sum_{i=1}^M \mathbf{P}_{ai}$ and $\bar{\mathbf{P}}_q = \frac{1}{M}\sum_{i=1}^M \mathbf{P}_{qi}$ be the point-cloud centroids, and let $\tilde{\mathbf{P}}_{ai} = \mathbf{P}_{ai} - \bar{\mathbf{P}}_a$, $\tilde{\mathbf{P}}_{qi} = \mathbf{P}_{qi} - \bar{\mathbf{P}}_q$ denote the centroid-centered point sets. The cross-covariance matrix and its SVD are:

$$H = \sum_{i=1}^{M} \tilde{\mathbf{P}}_{ai}\tilde{\mathbf{P}}_{qi}^\top, \quad H = U\Sigma V^\top, \tag{4}$$

$$\mathbf{R} = V \cdot \mathrm{diag}(1, 1, \det(VU^\top)) \cdot U^\top, \quad \mathbf{t} = \bar{\mathbf{P}}_q - \mathbf{R}\bar{\mathbf{P}}_a. \tag{5}$$

Geometric Self-Verification Solver (GVS). The closed-form SVD solution in Eq. equation 5 is a least-squares fit over all retained correspondences and is sensitive to residual outliers that survive mask filtering. Our key insight is that this initial fit, even when perturbed, is accurate enough to expose geometric inconsistencies. GVS uses the current pose estimate to verify correspondence consistency and re-solves the pose in closed form. Given the pose $(\mathbf{R}^{(k)}, \mathbf{t}^{(k)})$ at iteration k, we compute the 3D alignment residual of each correspondence:

$$e_i^{(k)} = \|\mathbf{R}^{(k)}\mathbf{P}_{ai} + \mathbf{t}^{(k)} - \mathbf{P}_{qi}\|_2. \tag{6}$$

We then retain the fraction γ of correspondences with the smallest residuals,

$$\mathcal{C}^{(k+1)} = \left\{ i \in \mathcal{C}^{(k)} \;\middle|\; e_i^{(k)} \le e^{(k)}_{(\lceil\gamma|\mathcal{C}^{(k)}|\rceil)} \right\}, \tag{7}$$

where $\mathcal{C}^{(0)}$ is the index set of $\mathcal{C}_{\text{geo}}$, $e^{(k)}_{(\lceil\gamma|\mathcal{C}^{(k)}|\rceil)}$ denotes the $\lceil\gamma|\mathcal{C}^{(k)}|\rceil$-th smallest residual and ties are broken by index so that exactly $\lceil\gamma|\mathcal{C}^{(k)}|\rceil$ correspondences survive; we then re-solve Eq. equation 4–equation 5 on the trimmed set. This loop is repeated for K_{iter} iterations (default $K_{\text{iter}} = 1$, $\gamma = 0.7$). If fewer than three correspondences remain, we fall back to the previous iteration's pose. Unlike iterative closest point variants relying on nearest-neighbor re-assignment, GVS operates on putative cross-view matches, requiring no pose initialization. The trimmed least-squares objective and detailed comparisons are provided in Appendix B and H.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on YCB-Video Xiang et al. (2018), NOCS-REAL275 Wang et al. (2019b), Toyota Light Hodaň et al. (2018), and LineMOD Hinterstoisser et al. (2012), spanning cluttered multi-object, category-level, and texture-poor regimes.

Protocol. We follow the anchor–query relative-pose protocol of Corsetti et al. (2024; 2026): each dataset contributes 2,000 anchor–query pairs (seed 42, shared by all methods), scored by a single evaluation code with pair-weighted aggregation; pairs with no pose count as zero.

Baselines. We compare against LatentFusion Park et al. (2020), OnePoseviaGen Geng et al. (2025) (CAD-free), GigaPose Nguyen et al. (2024) (CAD-based), PoseDiffusion Wang et al. (2023) and RelPose++ Lin et al. (2024a) (sparse-view), and Oryon Corsetti et al. (2024) plus Horyon Corsetti et al. (2026) (open-vocabulary), all re-run under identical protocols except Horyon (quoted). Privileged conditions (GT crop, GT scale, CAD models) are disclosed in Appendix D.

Metrics. We report ADD(S)-0.1d Hinterstoisser et al. (2012) (correct if average 3D distance $< 10\%$ of object diameter; ADD-S for symmetric objects) and BOP AR Hodaň et al. (2018) (mean of

Table 2: **Quantitative 6D pose estimation performance across four benchmarks.** Evaluated on 2,000 anchor–query pairs per dataset under our unified protocol (seed 42). Methods are categorized by reference regime; "Oracle" (in gray) denotes ground-truth mask upper bounds. The *Prior* column specifies the 2D mask condition: *Ours* denotes masks predicted by our open-vocabulary localization stage, and *Oracle* denotes ground-truth masks. Ground2Pose (light blue) achieves state-of-the-art accuracy under realistic predicted masks, with best practical results in **bold**. Δ denotes absolute improvement over the best competing practical baseline; the Average column compares overall means. Differences are computed from unrounded values. All ADD(S)-0.1d and AR values are percentages. †Evaluated under sparse-view protocol with GT crop and translation scale; ‡Quoted from Corsetti et al. (2026). Detailed baseline disclosures and failure breakdowns are provided in Sec. 4 and D.

Method	Prior	YCBV		NOCS		TOYL		LM		Average	
		ADD	AR	ADD	AR	ADD	AR	ADD	AR	ADD	AR
Sparse-view relative pose											
PoseDiffusion† (2023)	Oracle	4.3	3.3	0.3	0.9	1.0	1.0	0.2	0.8	1.5	1.5
RelPose++† (2024a)	Oracle	4.4	3.4	0.1	0.2	0.2	1.4	0.5	0.7	1.3	1.4
CAD-free											
LatentFusion (2020)	Ours	24.4	6.8	18.7	31.8	54.6	26.1	22.8	8.7	30.1	18.4
LatentFusion	Oracle	19.8	7.5	17.7	34.5	49.6	26.5	36.2	12.5	30.8	20.3
OnePoseviaGen (2025)	Ours	11.2	6.0	17.2	27.8	36.5	21.3	12.1	9.0	19.3	16.0
OnePoseviaGen	Oracle	11.6	6.0	10.5	22.2	32.0	21.5	13.9	11.5	17.0	15.3
CAD-based											
GigaPose (2024)	Ours	26.0	19.4	19.9	34.8	36.2	34.5	15.2	17.2	24.3	26.5
GigaPose	Oracle	25.5	21.1	17.1	34.7	39.2	35.2	21.3	22.9	25.8	28.5
Open-vocabulary											
Oryon (2024)	Ours	11.4	16.4	23.3	34.2	22.9	31.0	7.2	9.0	16.2	22.7
Oryon	Oracle	22.7	29.5	35.1	45.6	24.1	33.5	15.0	21.6	24.5	32.6
Horyon‡ (2026)	GDino+CATSeg	9.3	17.9	39.6	47.4	23.5	32.7	17.4	19.6	22.5	29.4
Horyon‡	GDino+Horyon	12.3	20.6	51.6	57.9	25.1	33.0	20.4	22.0	27.4	33.4
Ground2Pose	Ours	**58.5**	**37.6**	**86.2**	**63.8**	**83.2**	**41.2**	**55.6**	**26.3**	**70.9**	**42.2**
Ground2Pose	Oracle	86.0	52.9	92.3	69.1	84.3	41.6	88.1	39.6	87.7	50.8
Δ		+32.5	+17.0	+34.6	+5.9	+28.6	+6.7	+32.8	+4.3	+40.8	+8.8

VSD/MSSD/MSPD recalls over 5%–50% thresholds; per-method values in Appendix Table 11), plus mask IoU.

Components. Default localization: Grounded-SAM Ren et al. (2024) (Grounding-DINO Liu et al. (2024) + SAM Kirillov et al. (2023)); dense matching: LoFTR Sun et al. (2021); GVS: $\gamma = 0.7$, $K_{\text{iter}} = 1$, $K_{\text{top}} = \text{all}$ (fixed across datasets). All experiments on a single NVIDIA RTX 3090 with seed 42; significance via paired Wilcoxon tests and 10,000-sample bootstrap CIs. Full configuration details in Appendix D.

4.2 Main Results

Overall Performance. Under the predicted-mask setting, Ground2Pose leads every baseline on all four benchmarks by wide margins: 86.2 vs. 51.6 on NOCS (vs. Horyon), 58.5 vs. 26.0 on YCBV (vs. GigaPose), and 55.6 vs. 22.8 on LineMOD (vs. LatentFusion). The smallest margin is 28.6 points; margins against LatentFusion and GigaPose are statistically significant (paired Wilcoxon, see below and Appendix G). With GT masks (oracle), Ground2Pose ranks first across all benchmarks (86.0/92.3/84.3/88.1 ADD(S)-0.1d), outperforming the nearest competitor by 34.7–60.5 points. Both Oryon and Horyon, our most direct competitors, trail by 34.6–62.9 points under the identical metric.

Modular Mask Sources, Dataset-wise Analysis, and CAD Comparison. Pose accuracy broadly tracks mask quality across segmenters (Table 3), though not monotonically: wrong-instance masks can still yield usable poses when they cover the target's dominant surface, and SAM's multi-mask selection favors geometric consistency over raw IoU, so e.g. OWL-ViT+SAM attains the highest YCBV mask IoU yet our localization attains the highest YCBV accuracy. The remaining gap to the GT-mask oracle (27.5 and 32.5 points on YCBV/LM) is concentrated on cluttered and texture-poor benchmarks, confirming that our mask-agnostic backend composes freely with stronger segmenters.

Table 3: **Mask quality (IoU %) vs. pose accuracy (ADD(S)-0.1d).** Evaluated on four benchmarks with the geometric backend fixed (GVS on) while varying the localization front-end. Default Grounded-SAM is highlighted in blue; Oracle denotes ground-truth masks.

Mask source	Mask IoU				ADD(S)-0.1d			
	YCBV	NOCS	TOYL	LM	YCBV	NOCS	TOYL	LM
Talk2DINO (2025)	33.4	60.4	57.2	28.6	41.9	61.3	69.3	35.0
CLIPSeg (2022)	38.4	65.2	62.4	22.1	42.8	68.6	66.4	20.0
OWL-ViT+SAM (2022; 2023)	61.1	84.7	87.2	42.5	57.4	81.0	80.2	34.8
Ours	43.8	87.6	89.5	61.2	**58.5**	**86.2**	**83.2**	**55.6**
GT mask (oracle)	100	100	100	100	86.0	92.3	84.3	88.1

The nearest competitor differs by benchmark type (CAD-based on cluttered, trained fusion on clean), yet Ground2Pose still leads by at least 28.6 points on both. The CAD-based GigaPose trails Ground2Pose in every cell despite receiving privileged per-query detections and CAD models, demonstrating that a training-free, CAD-free framework can match or exceed template-based methods under our single-anchor, text-only setting.

Statistical Significance. The GVS-enabled model is significantly better on all four datasets under paired Wilcoxon signed-rank tests ($p < 10^{-6}$), with 95% bootstrap CIs of the gains excluding zero. Ground2Pose beats LatentFusion and GigaPose on all four datasets with per-pair advantages exceeding 28 points ($p < 10^{-75}$). Full statistics are in Appendix G; run-to-run std ≤ 0.09 points across seeds $\{42, 123, 7\}$.

4.3 Mask Quality Analysis

Table 3 evaluates the impact of localization quality on downstream pose estimation. Across 20 mask-source–dataset combinations, mask IoU and ADD(S)-0.1d exhibit a strong rank correlation ($\rho = 0.934$, Spearman), demonstrating that upstream segmentation sets the empirical accuracy ceiling. Default Grounded-SAM attains the highest pose accuracy across all four benchmarks. The substantial remaining gap to the GT-mask oracle (87.7% vs. 70.9% average) confirms that downstream performance is primarily bounded by upstream localization errors, which are dominated by instance confusion rather than boundary degradation (Appendix I).

4.4 Ablation Studies

Contribution of GVS. As shown in Table 4, under default Grounded-SAM localization, GVS improves ADD(S)-0.1d by $+2.80$ to $+4.25$ points across all four benchmarks (all $p < 10^{-6}$; Table 12). Under the detector-free Talk2DINO variant with noisier masks, GVS provides larger improvements of $+10.40$ to $+17.35$ points, demonstrating strong robustness to imperfect upstream segmentation.

GVS vs. robust estimators. Table 4 compares solvers under matched correspondences and outlier budgets (Appendix H). RANSAC-Kabsch trails GVS by 0.9–3.6 points even at 256 iterations due to sampling sensitivity on structured boundary errors. The certifiable TEASER++ falls below untrimmed Kabsch (trailing GVS by 17.1–35.1 points) because surviving outliers form spatially coherent boundary clusters rather than random gross errors, violating its assumption that inliers constitute the dominant clique. SC2-PCR is more robust to structured noise but still trails GVS by 6.1–14.7 points while requiring $\sim 34\times$ more solve time (5.4 ms vs. 0.16 ms). Deterministic trimming on full-set SVD is thus both the most accurate and fastest solver under this correspondence regime.

Interface Analysis across Mask Regimes. Controlled mask experiments clarify the decoupled roles of localization and geometric verification (full two-factor analysis in Appendix H, Table 16). Under ground-truth oracle masks, the accuracy ceiling rises to 87.7% on average, and the gap between untrimmed Kabsch and GVS narrows to just 3.3 points (84.3 vs. 87.7). In contrast, under weak localization (Talk2DINO), the GVS improvement over untrimmed Kabsch expands significantly to $+12.6$ points (39.3 vs. 51.9, Table 4). This confirms that while upstream localization establishes the empirical accuracy ceiling (spanning ~ 36 points across mask sources), correspondence-level geometric verification determines how closely the solver approaches that ceiling, providing the greatest benefit when localization is imperfect.

Table 4: Component ablation and geometric solver comparison under identical correspondences: ADD(S)-0.1d on all four benchmarks. GVS denotes the Geometric Self-Verification Solver (Sec. 3.3). Top block: geometric solver comparison under default localization; bottom block: weak-localization robustness under the detector-free Talk2DINO variant. All solvers in the top block receive identical LoFTR correspondences and mask filtering; robust-baseline outlier thresholds are matched to the GVS trimming budget ($\gamma = 0.7$ quantile aperture; full 2,000-pair protocol, Appendix H). All GVS gains are statistically significant under paired Wilcoxon signed-rank tests ($p < 10^{-6}$; Table 12). GVS is also the fastest robust solver at these correspondence scales (0.16 ms vs. 0.45–21.9 ms per solve; Appendix H).

Geometric Solver / Configuration	YCBV	NOCS	TOYL	LM
Default localization				
Plain Kabsch (w/o GVS, untrimmed)	54.25	82.65	79.65	52.80
+ RANSAC-Kabsch (32 it.)	57.00	82.60	80.20	52.70
+ RANSAC-Kabsch (256 it.)	57.60	82.60	80.80	53.20
+ TEASER++ (GNC)	41.45	51.10	53.90	25.60
+ SC2-PCR	50.70	72.45	68.50	49.50
Ground2Pose + GVS	**58.50**	**86.20**	**83.20**	**55.60**
GVS gain	*+4.25*	*+3.55*	*+3.55*	*+2.80*
Detector-free variant (Talk2DINO, weak localization)				
Variant (w/o GVS)	30.15	43.90	58.90	24.10
Variant + GVS	**41.90**	**61.25**	**69.30**	**35.00**
GVS gain	*+11.75*	*+17.35*	*+10.40*	*+10.90*

Table 5: **Computational cost of all methods (RTX 3090, per-pair).** All methods share the detection–segmentation front-end (+0.74 s/pair); times are end-to-end means over 2,000-pair runs. §Talk2DINO skips the front-end. †Sparse-view methods use GT crops/scale.

Method	Time (s)	VRAM (GB)	Training	CAD
Ground2Pose (default)	1.03	8.5	None	No
Ground2Pose (variant)§	**0.37**	8.1	None	No
GigaPose	0.80	15.0	None	Yes
LatentFusion	5.71	3.4	Trained	No
OnePoseviaGen	2.93	22.3	Trained	No
Oryon	0.97	18.8	Trained	No
PoseDiffusion†	1.00	0.7	Trained	No
RelPose++†	1.39	4.8	Trained	No

Hyperparameters and Design Principles. Diagnostic sweeps across benchmarks (Appendix H, Tables 14–15) confirm that GVS is robust to parameter choices: ADD(S)-0.1d remains within a 1.1-point band across $\gamma \in [0.5, 0.9]$, and $K_{\text{iter}} = 1$ with $K_{\text{top}} =$ all achieves the optimal accuracy-efficiency balance without per-dataset tuning. Furthermore, extensive probes show that while soft semantic reweighting marginally aids open-loop Kabsch, it degrades the full GVS pipeline (-1.30 YCBV, -2.05 LM; Appendix C.2), validating our core design principle: *robustness is most reliably achieved through hard 3D residual verification rather than soft heuristic reweighting*.

4.5 Qualitative Results and Failure Modes

Fig. 3 shows representative pairs where Ground2Pose succeeds while trained or CAD-based baselines fail; the margins in Table 2 hold at the level of individual frames. Qualitative pose predictions are in Appendix J; failure mode analysis and per-object breakdowns in Appendix E; qualitative mask comparisons in Appendix K. Scenario analysis and the verification-over-reweighting design principle (eight pre-registered negative probes) are detailed in Appendix C.

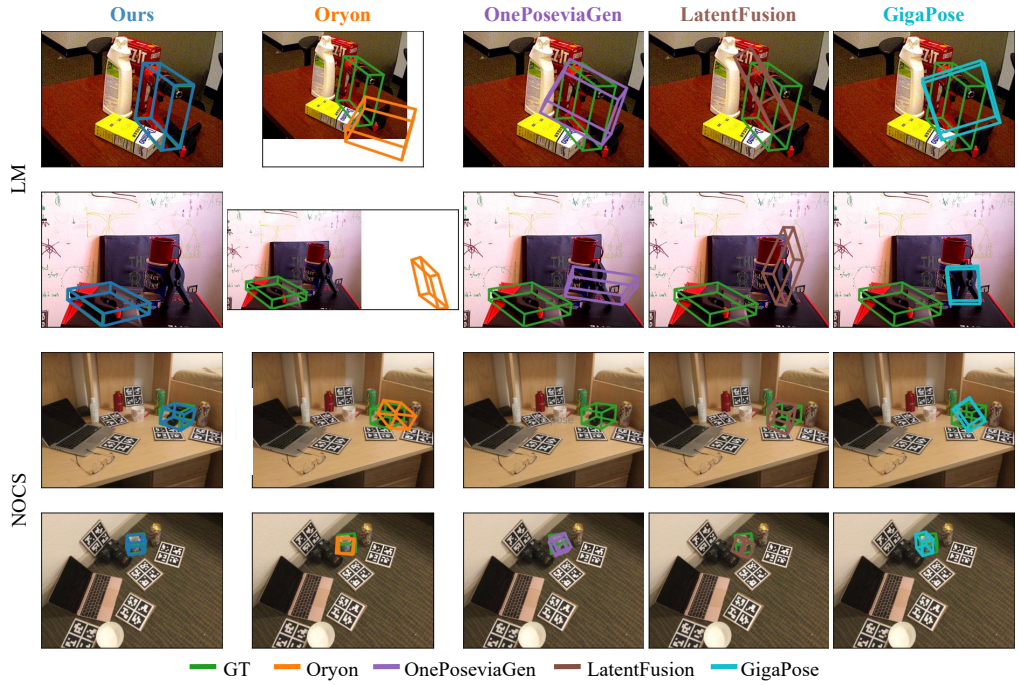


Figure 3: **Where Ground2Pose succeeds and baselines fail.** Predicted 3D bounding boxes overlaid on query images for pairs where Ground2Pose succeeds (ADD ≤ 0.1d) while at least two baselines fail, selected deterministically on seed-42 pairs from YCBV (top two rows) and NOCS-REAL275 (bottom two rows). Columns: Ground2Pose and four baselines (Oryon, OnePoseviaGen, LatentFusion, GigaPose); blue = predicted, green = ground truth; "no pose" marks methods returning no output. Full quantitative comparison in Table 2.

4.6 Computational Efficiency

Table 5 reports per-pair latency with every method carrying the same detection–segmentation front-end (0.74 s/pair). Ground2Pose costs 1.03 s/pair—within $\sim$1.3$\times$ of the lightest baselines and 5.5$\times$ faster than LatentFusion; the detector-free variant costs only 0.37 s. GVS adds negligible latency (one extra SVD). Unlike trained baselines, Ground2Pose requires no task-specific training and no CAD models.

5 Conclusion

We presented Ground2Pose, a training-free framework that estimates the 6D pose of novel objects from an anchor–query RGB-D view pair and a text description—no CAD model, no multi-view reference, and no task-specific training. The framework rests on two design claims: the localization stage sets an empirical accuracy ceiling (bounded from above by GT-mask interventions rather than by correlation alone; the rank correlation at $\rho = 0.934$ is its observational view), and within that bound, deterministic correspondence-level verification (GVS) recovers more accuracy than stochastic or certifiable alternatives while adding negligible compute cost. Eight pre-registered negative results generalize this into a "verification-over-reweighting" principle. Ground2Pose leads every baseline on every benchmark (58.5/86.2/83.2/55.6 ADD(S)-0.1d; $p < 10^{-75}$ against LatentFusion and GigaPose). All components are frozen pre-trained models; experiments are reproducible with fixed seeds (std ≤ 0.1).

Limitations. Localization remains the dominant error source in cluttered or texture-poor scenes (27.5 and 32.5 points to oracle on YCBV/LM); the anchor frame is assumed to come with its pose; multi-instance disambiguation requires additional spatial referring modules. Future work will explore higher-resolution mask prediction, multi-view fusion, and real-time deployment.

AI USE STATEMENT

Generative AI tools were used for partial translation and language editing to improve the readability, clarity, and grammatical consistency of the manuscript, as well as to assist in writing and editing experimental code. They were not used to generate data, formulate hypotheses or mathematical claims, design the methodology or experiments, analyze data, interpret results, or draw scientific conclusions. The authors reviewed all AI-assisted manuscript edits and verified that they preserved the intended technical meaning and did not introduce unsupported claims, results, or references. All AI-assisted code was inspected, tested, and validated by the authors for correctness before use. The authors take full responsibility for the final content of this work, including all text, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We conduct experiments on four publicly available datasets: YCB-Video Xiang et al. (2018), NOCS-REAL275 Wang et al. (2019b), Toyota Light Hodaň et al. (2018), and LineMOD Hinterstoisser et al. (2012). YCB-Video, Toyota Light, and LineMOD are available through the BOP benchmark at https://bop.felk.cvut.cz/datasets/, and NOCS-REAL275 is available at https://github.com/hughw19/NOCS_CVPR2019. Code will be made publicly available upon acceptance.

REFERENCES

Abdulrahman Alsumeri, Di-Hua Zhai, and Yuanqing Xia. FreePose: Zero-shot 6D object pose estimation using pretrained foundation models. *IEEE Transactions on Circuits and Systems for Video Technology*, 36:5038–5052, 2025. doi: 10.1109/TCSVT.2025.3627799.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22025–22035, 2025. doi: 10.1109/ICCV51701.2025.02045.

Dingding Cai, Janne Heikkilä, and Esa Rahtu. OVE6D: Object viewpoint encoding for depth-based 6D object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6793–6803, 2022. doi: 10.1109/CVPR52688.2022.00668.

Andrea Caraffa, Davide Boscaini, Amir Hamza, and Fabio Poiesi. FreeZe: Training-free zero-shot 6D pose estimation with geometric and vision foundation models. In *European Conference on Computer Vision (ECCV)*, pp. 414–431, 2024. doi: 10.1007/978-3-031-73226-3_24.

Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. SC2-PCR: A second order spatial compatibility for efficient and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13211–13221, 2022. doi: 10.1109/CVPR52688.2022.01287.

Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *International Conference on Pattern Recognition (ICPR)*, pp. 545–548, 2002. doi: 10.1109/ICPR.2002.1047997.

Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. Open-vocabulary object 6D pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18071–18080, 2024. doi: 10.1109/CVPR52733.2024.01711.

Jaime Corsetti, Davide Boscaini, Francesco Giuliari, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. High-resolution open-vocabulary object 6D pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):2066–2077, 2026. doi: 10.1109/TPAMI.2025.3624589.

Weijian Deng, Dylan Campbell, Chunyi Sun, Jiahao Zhang, Shubham Kanitkar, Matthew E. Shaffer, and Stephen Gould. Pos3R: 6D pose estimation for unseen objects made easy. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16818–16828, 2025. doi: 10.1109/CVPR52734.2025.01567.

Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. PoseRBPF: A rao-blackwellized particle filter for 6-D object pose tracking. *IEEE Transactions on Robotics*, 37 (5):1328–1342, 2021. doi: 10.1109/TRO.2021.3056043.

Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 998–1005, 2010. doi: 10.1109/CVPR.2010.5540108.

Guowei Du, Kai Wang, Shiguo Lian, and Kai Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review. *Artificial Intelligence Review*, 54:1677–1734, 2021. doi: 10.1007/s10462-020-09888-5.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981. doi: 10.1145/358669.358692.

Zheng Geng, Nan Wang, Shaocong Xu, Chongjie Ye, Bohan Li, Zhaoxi Chen, Sida Peng, and Hao Zhao. One view, many worlds: Single-image to 3D object meets generative domain randomization for one-shot 6D pose estimation. In *Conference on Robot Learning (CoRL)*, 2025.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. OnePose++: Keypoint-free one-shot object pose estimation without CAD models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 35184–35197, 2022. doi: 10.52202/068431-2544.

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 (CVPR)*, pp. 3002–3012, 2021. doi: 10.1109/CVPR46437.2021.00302.

Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In *Asian Conference on Computer Vision (ACCV)*, pp. 548–562, 2012. doi: 10.1007/978-3-642-37331-2_42.

Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. In *European Conference on Computer Vision (ECCV)*, pp. 19–35, 2018. doi: 10.1007/978-3-030-01249-6_2.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6D: Model-free 6D pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11633–11643, 2025. doi: 10.1109/CVPR52734.2025.01086.

Amy Lin, Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose++: Recovering 6D poses from sparse-view observations. In *International Conference on 3D Vision (3DV)*, pp. 106–115, 2024a. doi: 10.1109/3DV62453.2024.00126.

Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27906–27916, 2024b. doi: 10.1109/CVPR52733.2024.02636.

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. doi: 10.1007/s11263-025-02646-6.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images. In *European Conference on Computer Vision (ECCV)*, pp. 298–315, 2022. doi: 10.1007/978-3-031-19824-3_18.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7076–7086, 2022. doi: 10.1109/CVPR52688.2022.00695.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *European Conference on Computer Vision (ECCV)*, pp. 728–755, 2022. doi: 10.1007/978-3-031-20080-9_42.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9903–9913, 2024. doi: 10.1109/CVPR52733.2024.00945.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

Kiru Park, Timothy Patten, and Markus Vincze. Neural object learning for 6D pose estimation using a few cluttered images. In *European Conference on Computer Vision (ECCV)*, pp. 656–673, 2020. doi: 10.1007/978-3-030-58548-8_38.

Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-wise voting network for 6DoF pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4561–4570, 2019. doi: 10.1109/CVPR.2019.00469.

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 *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8922–8931, 2021. doi: 10.1109/CVPR46437.2021.00881.

Jiaming Sun, Zihao Wang, Siyuan Zhang, Yifan He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6825–6834, 2022. doi: 10.1109/CVPR52688.2022.00670.

Philip H. S. Torr and Andrew Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000. doi: 10.1006/cviu.1999.0832.

Yuyang Tu, Yunlong Wang, Hui Zhang, Wenkai Chen, and Jianwei Zhang. Language-embedded 6D pose estimation for tool manipulation. *IEEE Robotics and Automation Letters*, 10(9):8618–8625, 2025. doi: 10.1109/LRA.2025.3587559.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Javier Villena Toro and Mehdi Tarkian. Geo6DPose: Fast zero-shot 6D object pose estimation via geometry-filtered feature matching. *arXiv preprint arXiv:2512.10674*, 2025.

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 (CVPR)*, pp. 3343–3352, 2019a. doi: 10.1109/CVPR.2019.00346.

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 (CVPR)*, pp. 2642–2651, 2019b. doi: 10.1109/CVPR.2019.00275.

Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9739–9749, 2023. doi: 10.1109/ICCV51070.2023.00896.

Bowen Wen, Chaitanya Mitash, Baoquan Chen, and Kostas E. 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 (IROS)*, pp. 8067–8074, 2021. doi: 10.1109/IROS51168.2021.9635991.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, 2024. doi: 10.1109/CVPR52733.2024.01692.

Chenrui Wu, Long Chen, Shenglong Wang, Han Yang, and Junjie Jiang. Geometric-aware dense matching network for 6D pose estimation of objects from RGB-D images. *Pattern Recognition*, 137:109293, 2023. doi: 10.1016/j.patcog.2022.109293.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. doi: 10.15607/RSS.2018.XIV.019.

Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox. Unseen object instance segmentation for robotic environments. *IEEE Transactions on Robotics*, 37(5):1343–1359, 2021. doi: 10.1109/TRO.2021.3060341.

Heng Yang, Jingnan Shi, and Luca Carlone. TEASER: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2021. doi: 10.1109/TRO.2020.3033695.

Andy Zeng, Kuan-Ting Yu, Shuran Wang, Jeffrey Ichnowski, Menglong Peng, Igor Mordatch, and Alberto Rodriguez. Multi-view self-supervised deep learning for 6D pose estimation in the amazon picking challenge. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1386–1393, 2017. doi: 10.1109/ICRA.2017.7989165.

Guangliang Zhou, Yi Yan, Deming Wang, and Qijun Chen. A novel depth and color feature fusion framework for 6D object pose estimation. *IEEE Transactions on Multimedia*, 23:1630–1639, 2021. doi: 10.1109/TMM.2020.3001533.

APPENDIX

The appendix is organized as follows. Appendix A expands the related work, Appendix B provides additional method derivations, Appendix C discusses scenario-dependent behavior and the "verification-over-reweighting" principle, and Appendix D documents the full configuration, evaluation protocol, and failure cases. Quantitative evidence is extended through per-object results in Appendix E, full BOP metrics in Appendix F, and per-pair statistical analyses in Appendix G. Appendix H reports hyperparameter, solver, and pre-registered negative-result probes, while Appendix I diagnoses the localization–geometry interface. Finally, Appendix J and Appendix K provide qualitative pose and mask visualizations, respectively.

A EXTENDED RELATED WORK

A.1 TRAINING-FREE 6D POSE ESTIMATION

We expand on the paradigms discussed in Sec. 2. CAD-reliant methods Drost et al. (2010); Caraffa et al. (2024); Villena Toro & Tarkian (2025) such as PPF Drost et al. (2010), FreeZe Caraffa et al. (2024), and Geo6DPose Villena Toro & Tarkian (2025) eliminate pose-specific training by matching against 3D CAD meshes or rendered templates. Reference-based methods Liu et al. (2022); Sun et al. (2022); Deng et al. (2025); Lee et al. (2025) include Gen6D Liu et al. (2022), OnePose Sun et al. (2022)/OnePose++ He et al. (2022) (requiring multi-view sequences or SfM maps), BundleTrack Wen et al. (2021), FreePose Alsumeri et al. (2025) (CAD-free but no language), and Any6D Lee et al. (2025) (pre-segmented category masks, no language instructions).

A.2 LANGUAGE-GUIDED POSE ESTIMATION

Language guidance enables querying novel objects via free-form text. Beyond Oryon Corsetti et al. (2024) and Horyon Corsetti et al. (2026) (discussed in Sec. 2), SAM-6D Lin et al. (2024b) explores language prompts alongside CAD models. The core limitations of task-specific supervision and open-loop error propagation motivate our decoupled, training-free architecture with GVS verification.

A.3 OPEN-VOCABULARY SEGMENTATION AND GEOMETRIC LOCALIZATION

Open-vocabulary detection and segmentation models serve as the perceptual front-end for language-guided pose estimation, fundamentally powered by cross-modal representation learning such as CLIP Radford et al. (2021) and self-supervised visual backbones such as DINOv2 Oquab et al. (2023). CLIPSeg Lüddecke & Ecker (2022) decodes vision-language representations into coarse text-conditioned masks, OWL-ViT Minderer et al. (2022) predicts open-vocabulary bounding boxes, and Grounded-SAM Ren et al. (2024) combines Grounding-DINO Liu et al. (2024) with SAM Kirillov et al. (2023) to produce detection-grounded instance masks.

Crucially, while these models excel at 2D semantic benchmarks, their *downstream 3D geometric utility* varies substantially. Standard 2D vision-language models optimize 2D semantic overlap rather than 3D boundary fidelity. Consequently, boundary leakage improperly introduces background surface depth, and wrong-instance ambiguity in cluttered scenes injects completely erroneous 3D coordinate structures—a persistent bottleneck also documented in robotic unseen object segmentation Xie et al. (2021); Zhou et al. (2021). Treating open-vocabulary localization as a modular front-end enables us to systematically diagnose this bottleneck between localization and geometry under a frozen geometric backend (Sec. 4). This approach further reveals that upstream semantic filtering alone cannot eliminate 3D geometric noise, which necessitates an explicit downstream verification mechanism.

A.4 ROBUST GEOMETRIC REGISTRATION AND VERIFICATION

Beyond RANSAC Fischler & Bolles (1981) (discussed in Sec. 2), MLESAC Torr & Zisserman (2000) similarly relies on random sampling. Modern estimators like SC2-PCR Chen et al. (2022) and TEASER++ Yang et al. (2021) offer theoretical robustness against extreme outlier ratios but incur computational overhead on dense fields; we include both as matched baselines in our solver

comparison (Table 4). Deterministic trimmed least-squares Chetverikov et al. (2002); Wu et al. (2023) iteratively discards high-residual pairs for fast convergence; GVS adopts this principle, using the full-set closed-form SVD as verifier and re-solving with negligible compute cost.

B Detailed Method Derivations

B.1 Top-K Correspondence Retention

The Top-K reduction (Sec. 3.2) is treated as an implementation choice rather than a contribution. Unless stated otherwise, all correspondences in $\mathcal{C}_{\text{geo}}$ are used ($K_{\text{top}} = \text{all}$, i.e., $\mathcal{C}_{\text{final}} = \mathcal{C}_{\text{geo}}$), and the effect of K_{top} is analyzed in the hyperparameter study (Appendix H).

B.2 Trimmed Least-Squares Objective and RANSAC Comparison

Mechanistically, each re-solve of GVS minimizes the trimmed least-squares objective over the retained set:

$$\min_{\mathbf{R},\mathbf{t}} \sum_{i \in \mathcal{C}^{(k+1)}} \|\mathbf{R}\mathbf{P}_{ai} + \mathbf{t} - \mathbf{P}_{qi}\|_2^2 , \tag{8}$$

which places GVS in the deterministic trimmed least-squares family Chetverikov et al. (2002), applied here to putative cross-view correspondences rather than nearest-neighbor pairs. Unlike RANSAC Fischler & Bolles (1981), which draws random minimal subsets, GVS is fully deterministic and initialized from the full correspondence set, with each iteration requiring only one additional closed-form SVD and γ held fixed at 0.7 across all datasets. Sec. 4 compares this directly against a RANSAC-Kabsch solver on identical correspondences.

C Discussion: Scenario Analysis and Design Principles

C.1 Discussion: Scenario Split, Multi-Object Scenes, and Anchor Selection

The dataset-wise pattern in Table 2 is consistent: under predicted masks Ground2Pose leads every baseline on every benchmark, with margins ranging from 28.6 points (vs. LatentFusion on TOYL) to 34.6 points over the strongest competitor Horyon on NOCS (and 62.9 points vs. Oryon); margins against LatentFusion and GigaPose are statistically significant (Sec. 4.2). The reasons are structural. On YCBV and NOCS, heavy clutter, occlusion, and symmetric instances break template and reference-image matching, while our dense RGB-D matching with geometric self-verification tolerates partial visibility—yielding the largest margins; on LM and TOYL, each scene contains one cleanly visible object on a plain background, where the trained fusion baseline LatentFusion is at its strongest, yet Ground2Pose still leads by 28.6–32.8 points. Both results are useful for deployment: fusion and template pipelines remain competitive when their assumptions hold, whereas training-free RGB-D pipelines like Ground2Pose are preferable in open-world, cluttered manipulation settings where no CAD, no training, and no clean reference view can be assumed. Our current formulation assumes a single target object specified by the text prompt. In scenes with multiple instances of the same category (e.g., two red bowls), the text prompt alone cannot disambiguate: our default configuration already employs Grounding-DINO + SAM but keeps only the highest-confidence detection, so it cannot distinguish which instance the user intends. Spatial referring expressions (e.g., "the bowl on the left") combined with a referring-aware ranking of the detected instances are a natural extension—we leave this to future work. Regarding anchor selection, our pipeline assumes a given anchor image with the target object clearly visible. In practice, anchor quality affects matching coverage: frontal, unoccluded views yield the best correspondences, while extreme viewpoint changes ($>60°$) degrade LoFTR's matching density. Automatic anchor selection via view scoring Sun et al. (2022) is a promising direction.

C.2 Design Principle: Verification over Reweighting

Pose-to-mask closed-loop feedback. A natural extension of the GVS is to close the loop at the segmentation level: after the first pose estimate, the predicted pose can re-prompt the segmenter (e.g., by projecting the anchor mask into the query view) to correct localization errors. A pilot

study on YCBV and LineMOD found that this pose-to-mask feedback changes ADD(S)-0.1d by $-1.8/-3.4$ points respectively—failing our pre-registered adoption criterion of +1.0 on both datasets. The re-prompted masks leave the aggregate mask IoU essentially unchanged (Grounded-SAM masks are already near-saturated) while perturbing the correspondence set that the pose stage consumes, which degrades rather than improves the estimate. We therefore keep the single-pass localization as the default and view closed-loop segmentation as promising primarily for weaker localization front-ends.

Geometry-guided mask selection. SAM's multimask output provides three candidate masks ranked by predicted IoU. A natural question is whether downstream geometric consistency (inlier count after mask-constrained matching) is a better selection criterion than the task-agnostic predicted IoU. We evaluate this on 100 YCBV pairs: replacing predicted-IoU selection with geometric ranking yields identical ADD(S)-0.1d (58.0% vs. 58.0%), confirming that SAM's internal quality metric is already well-aligned with downstream pose accuracy. This further supports a diminishing-returns observation: once a strong segmenter is chosen, additional geometric refinement at the mask level provides diminishing returns.

Boundary-distance soft weights. A complementary probe asks whether *soft* mask-quality information can replace the GVS's hard residual trimming: we re-weighted each correspondence by the mask boundary distance at its two endpoints ($w_i = (1 - e^{-d_i^2/2\sigma^2})$, $\sigma = 5\,\text{px}$, routed through the same weighted-Kabsch machinery). On 100 YCBV pairs with oracle masks eroded by radius $r \in \{0, 4, 8, 12\}$ px, the soft weights lose 3–10 points relative to the unweighted GVS baseline, up to 10 points at the strongest erosion. The reason is informative: silhouette points are geometrically the *most* informative for rotation, so down-weighting them removes exactly the correspondences that carry the most pose information; erosion damage is better treated as outliers than as low-confidence inliers—which is precisely what the GVS's residual trimming does (it discards points because they disagree geometrically, independent of their position in the mask).

Collecting the negative results of this paper yields a consistent design principle for composing segmentation and geometric reasoning in pose estimation: robustness must come from *hard geometric verification*, not from *soft semantic or mask-quality reweighting*. Across eight pre-registered probes—semantic correspondence reweighting (Sec. 4.4), boundary-distance soft weights (Sec. C.2), pose-to-mask closed-loop feedback, geometry-guided SAM mask selection, top-k instance rescue, adaptive trimming fraction, SAM multimask arbitration, and attention-guided mask refinement (Appendix H)—every attempt to improve pose through soft semantic or mask-quality reweighting, or through coarser-grained verification, failed to pass our pre-registered adoption criterion, while the only reliable gain comes from per-correspondence 3D residual verification (the GVS, Sec. 3.3). The common thread is a hierarchy of granularity: geometric verification is reliable exactly where it has complete information—per-correspondence 3D residuals—and degrades monotonically as the decision moves to coarser objects (mask quality, instance identity, trimming fraction) for which no discriminative training-free signal exists. We therefore recommend correspondence-level geometric verification as the default mechanism for injecting robustness into training-free pose pipelines, and report these results to calibrate expectations for reweighting-based extensions.

D Method Details and Disclosures

Full Configuration. The default localization stage is Grounded-SAM Ren et al. (2024) (Grounding-DINO Liu et al. (2024) `grounding-dino-base` boxes prompting SAM Kirillov et al. (2023) sam-vit-base), with text prompts lowercased and terminated with a period following the Grounding-DINO convention; we use `box_threshold`$= 0.25$ and `text_threshold`$= 0.2$, keep the highest-scoring box, and prompt SAM in multi-mask mode, selecting among its three candidate masks the one with the highest predicted-IoU score. If no box passes the thresholds, we retry once with both thresholds set to zero and keep the top-scoring detection, so the pipeline always receives a mask. The alternative mask sources of Table 3 use publicly released weights without any fine-tuning: CLIPSeg (clipseg-rd64-refined) Lüddecke & Ecker (2022) and OWL-ViT (owlvit-base-patch32) Minderer et al. (2022), whose detected box prompts the same SAM. Each variant replaces only the segmentation stage; all remaining stages and hyperparameters are unchanged. For dense matching we use LoFTR Sun et al. (2021) with outdoor pre-trained weights in evaluation mode and its default confidence threshold. GVS hyperparameters are fixed across datasets (no per-dataset

Table 6: Failed pairs (no pose returned) per method and dataset under predicted-mask (*Ours*) and GT-mask (*Oracle*, in gray) conditions (the *Prior* column distinguishes the 2D mask condition); failures count as zero in all metrics.

Method	Prior	YCBV	NOCS	TOYL	LM
Ground2Pose	Ours	0	0	0	0
Ground2Pose	Oracle	0	0	0	0
LatentFusion	Ours	0	0	0	0
LatentFusion	Oracle	2	0	0	0
OnePoseviaGen	Ours	120	2	175	5
OnePoseviaGen	Oracle	142	2	161	1
GigaPose	Ours	0	0	0	0
GigaPose	Oracle	0	0	0	0
Oryon	Ours	20	0	26	27
Oryon	Oracle	8	0	5	0

tuning): $\gamma = 0.7$ (keep ratio, Eq. equation 7), $K_{\text{iter}} = 1$ (iterations), and all correspondences in $\mathcal{C}_{\text{geo}}$ are used ($K_{\text{top}} = $ all; we write K_{top} to avoid confusion with the camera intrinsics $\mathbf{K}$). All experiments use the sensor depth maps provided by the datasets, on a single NVIDIA RTX 3090 (24GB). Reported numbers use the shared fixed seed (42); the full configuration is additionally repeated with seeds $\{123, 7\}$ on all four datasets. For significance testing we record per-pair ADD(S) outcomes at seed 42 and run paired Wilcoxon signed-rank tests (base vs. full on identical pairs) together with 10,000-sample bootstrap 95% confidence intervals over pairs (SciPy).

Sparse-view protocol. Following Corsetti et al. (2026), PoseDiffusion and RelPose++ receive RGB-only input, the ground-truth query crop (cropped from the ground-truth mask), and the ground-truth translation magnitude, which converts their relative predictions into metric translations; they use no depth, no predicted masks, and no CAD models.

Failure disclosure. Pairs where a method returns no pose count as zero in every table. OnePoseviaGen suffers occasional OOM/process failures (Table 6); the remaining methods fail on at most 27 of 2000 pairs (Oryon), typically fewer than 15.

Oryon. Re-run with its official checkpoint under both its predicted-mask and GT-mask conditions on our exact 2000-pair lists; per-pair outputs are aggregated with the same pair weighting as all other methods.

Failure Case Analysis. We identify three primary failure modes: (1) *Cluttered multi-object scenes* (YCBV): text similarity responds to distractor objects of similar appearance, producing masks that cover the wrong instance. A per-pair probe (192 pairs per dataset, seed 42) shows that this wrong-instance mode dominates mask failures: mask IoU falls below 0.5 on 40.6% of YCBV and 32.8% of LineMOD pairs, whereas boundary erosion accounts for a much smaller share—consistent with Fig. 4b, where Grounded-SAM sits below the erosion curve precisely because its errors are wrong-instance selections rather than boundary damage. (2) *Small objects at low resolution* (LineMOD): small objects cover few pixels, and open-vocabulary segmenters miss them or bleed into the background, explaining the low mask IoU on this dataset. (3) *Out-of-vocabulary objects*: for categories distant from the segmenter's training vocabulary (Grounding-DINO/SAM training corpora), the detection–segmentation pipeline produces less reliable masks, degrading mask quality. The per-object breakdown in Appendix E shows predominantly positive but instance-level heterogeneous effects: on YCB-Video, 15 of 21 objects improve and none regress, with large gains on challenging items (e.g., $+11.4$ on obj 14, $+17.5$ on obj 15, $+12.5$ on obj 21; Table 7); on NOCS-REAL275, 6 of 13 objects improve and 4 regress (Table 8), the largest regression being -13.1 on the category-level symmetric obj 3; on Toyota Light, 12 of 21 improve and 5 regress (Table 9); and on LineMOD, 12 of 15 improve with only two negligible regressions (≤ 0.7 points; Table 10). Pair-weighted accuracy improves on every benchmark, while the regressing objects concentrate on category-level symmetric instances where trimming also removes weakly-textured surface correspondences. Remaining errors concentrate on a few small or texture-poor objects: on YCBV, objects 9 and 3 stay at 9.2% and 18.0%, and on LineMOD objects 12 and 2 at 1.9% and 8.8%—consistent with failure modes (1) and (2); crucially, GVS improves or preserves these hardest objects rather than sacrificing them.

E PER-OBJECT QUANTITATIVE RESULTS

Tables 7–10 report the per-object ADD(S)-0.1d success rates of our default configuration on all four benchmarks, aggregated from the per-pair evaluation outputs of seed 42 (n=2000 pairs per dataset, pair-weighted); per-row differences Δ are computed from unrounded success rates. GVS delivers large gains on many instances across asymmetric, textured, and cluttered objects (e.g., $+17.5$ on YCBV obj 15, $+22.7$ on NOCS obj 13, $+25.8$ on TOYL obj 17) by successfully removing background plane contamination, while preserving baseline accuracy on most challenging texture-poor objects; per-object outcomes overall (improve/regress/tie) are 15/0/6 on YCBV, 6/4/3 on NOCS, 12/5/4 on TOYL, and 12/2/1 on LineMOD (with LM regressions ≤ 0.7 points).

Table 7: Per-object ADD(S)-0.1d success rate on YCB-Video under default Grounded-SAM localization, without (w/o GVS) and with (w/ GVS) the geometric self-verification solver; n = evaluation pairs per object (2,000 total, seed 42, pair-weighted). 15 of 21 objects improve, 0 regress.

Objects 1–11					*Objects 12–21*				
Object	n	w/o GVS	w/ GVS	Δ	Object	n	w/o GVS	w/ GVS	Δ
1	159	57.9	58.5	+0.6	12	177	62.2	66.1	+4.0
2	107	61.7	66.4	+4.7	13	45	51.1	55.6	+4.4
3	278	17.6	18.0	+0.4	14	35	88.6	100.0	+11.4
4	389	67.6	73.0	+5.4	15	171	71.4	88.9	+17.5
5	49	100.0	100.0	+0.0	16	18	72.2	72.2	+0.0
6	170	24.7	25.9	+1.2	17	5	100.0	100.0	+0.0
7	12	33.3	41.7	+8.3	18	43	95.3	95.3	+0.0
8	14	85.7	85.7	+0.0	19	32	31.2	34.4	+3.1
9	98	7.1	9.2	+2.0	20	51	35.3	39.2	+3.9
10	41	65.8	78.0	+12.2	21	8	37.5	50.0	+12.5
11	98	100.0	100.0	+0.0					

Table 8: Per-object ADD(S)-0.1d success rate on NOCS-REAL275 under default Grounded-SAM localization, without (w/o GVS) and with (w/ GVS) the geometric self-verification solver; n = evaluation pairs per object (2,000 total, seed 42, pair-weighted). 6 of 13 objects improve, 4 regress.

Objects 1–7					*Objects 8–13*				
Object	n	w/o GVS	w/ GVS	Δ	Object	n	w/o GVS	w/ GVS	Δ
1	13	100.0	100.0	+0.0	8	170	97.1	95.9	−1.2
2	150	64.7	60.7	−4.0	9	285	100.0	100.0	+0.0
3	175	74.3	61.1	−13.1	10	196	99.0	98.0	−1.0
4	84	89.3	89.3	+0.0	11	106	42.4	65.1	+22.6
5	223	76.2	78.0	+1.8	12	207	88.4	92.8	+4.3
6	55	65.5	80.0	+14.5	13	247	70.9	93.5	+22.7
7	89	95.5	98.9	+3.4					

Table 9: Per-object ADD(S)-0.1d success rate on Toyota Light under default Grounded-SAM localization, without (w/o GVS) and with (w/ GVS) the geometric self-verification solver; n = evaluation pairs per object (2,000 total, seed 42, pair-weighted). 12 of 21 objects improve, 5 regress.

Objects 1–11					*Objects 12–21*				
Object	n	w/o GVS	w/ GVS	Δ	Object	n	w/o GVS	w/ GVS	Δ
1	107	72.0	69.2	−2.8	12	88	98.9	98.9	+0.0
2	63	98.4	98.4	+0.0	13	97	76.3	81.4	+5.1
3	129	77.5	81.4	+3.9	14	112	81.2	83.9	+2.7
4	126	92.9	92.9	+0.0	15	90	61.1	73.3	+12.2
5	55	94.5	90.9	−3.6	16	71	83.1	78.9	−4.2
6	82	74.4	76.8	+2.4	17	66	25.8	51.5	+25.8
7	92	66.3	70.7	+4.4	18	122	80.3	82.0	+1.6
8	106	71.7	87.7	+16.0	19	97	91.8	90.7	−1.0
9	77	89.6	89.6	+0.0	20	64	65.6	73.4	+7.8
10	148	96.6	94.6	−2.0	21	88	69.3	77.3	+8.0
11	120	85.0	89.2	+4.2					

Table 10: Per-object ADD(S)-0.1d success rate on LineMOD under default Grounded-SAM localization, without (w/o GVS) and with (w/ GVS) the geometric self-verification solver; n = evaluation pairs per object (2,000 total, seed 42, pair-weighted). 12 of 15 objects improve, 2 regress.

Objects 1–8					*Objects 9–15*				
Object	n	w/o GVS	w/ GVS	Δ	Object	n	w/o GVS	w/ GVS	Δ
1	103	15.5	21.4	+5.8	9	147	86.4	93.2	+6.8
2	147	6.1	8.8	+2.7	10	128	18.0	19.5	+1.6
3	137	95.6	94.9	−0.7	11	131	69.5	71.0	+1.5
4	139	77.0	78.4	+1.4	12	157	0.6	1.9	+1.3
5	114	99.1	99.1	+0.0	13	129	58.9	64.3	+5.4
6	120	85.8	88.3	+2.5	14	130	60.8	66.9	+6.1
7	138	54.4	53.6	−0.7	15	141	20.6	21.3	+0.7
8	139	54.7	62.6	+7.9					

F FULL BOP METRICS (VSD / MSSD / MSPD)

Table 11: **Full BOP recall metrics (VSD / MSSD / MSPD, pair-weighted) for all methods on the four benchmarks.** Each score is a recall value—the average recall over the 5%–50% threshold range (for VSD and MSSD relative to the object's diameter, for MSPD relative to the image size)—so *higher is better*, and Table 2 reports their arithmetic mean as the BOP average recall AR. 2,000 pairs, seed 42; failures count as zero. The *Prior* column specifies the 2D mask condition: *Ours* denotes masks predicted by our open-vocabulary localization stage, and *Oracle* denotes ground-truth masks. The practical (*Ours*) row of Ground2Pose is highlighted in light blue; best practical results in each column are shown in **bold**; oracle rows are shown in gray for reference only. Oryon is re-run under both conditions on our exact pairs; PoseDiffusion and RelPose++ follow the sparse-view protocol (GT crop, GT translation scale).

Method	Prior	YCBV			NOCS			TOYL			LM		
		VSD	MSSD	MSPD	VSD	MSSD	MSPD	VSD	MSSD	MSPD	VSD	MSSD	MSPD
Sparse-view relative pose													
PoseDiffusion†	Oracle	1.25	4.96	3.69	0.11	1.00	1.51	0.03	0.95	2.01	0.01	0.48	1.97
RelPose++†	Oracle	1.66	4.96	3.55	0.03	0.22	0.35	0.06	1.39	2.86	0.05	0.49	1.66
CAD-free													
LatentFusion	Ours	6.71	7.73	5.87	7.48	33.93	54.02	11.77	26.51	40.09	2.92	7.52	15.79
LatentFusion	Oracle	5.94	9.16	7.41	8.12	36.90	58.43	11.88	26.41	41.34	5.11	10.60	21.87
OnePoseviaGen	Ours	3.97	8.21	5.72	4.18	30.74	48.43	9.65	22.82	31.54	1.38	9.53	16.16
OnePoseviaGen	Oracle	2.41	9.01	6.49	2.68	19.15	44.92	7.64	23.96	32.99	1.54	11.30	21.71
CAD-based													
GigaPose	Ours	17.48	17.85	22.89	13.61	24.70	65.98	**20.67**	28.70	**54.14**	6.17	12.14	**33.25**
GigaPose	Oracle	18.21	18.60	26.49	12.84	23.28	68.04	21.32	29.58	54.56	8.23	16.18	44.25
Open-vocabulary													
Oryon	Ours	12.64	21.22	15.45	24.79	38.29	39.55	10.89	39.62	42.41	2.92	10.70	13.36
Oryon	Oracle	25.18	37.45	25.98	31.20	49.80	55.78	11.86	43.14	45.35	8.36	25.02	31.41
Ground2Pose	Ours	**37.96**	**40.39**	**34.48**	**49.34**	**74.09**	**67.88**	19.89	**50.35**	53.20	**19.36**	**28.73**	30.82
Ground2Pose	Oracle	56.39	55.14	47.12	53.84	79.54	74.00	20.43	50.71	53.57	30.03	42.68	46.01

G PER-PAIR STATISTICAL RIGOR

Per-pair Wilcoxon signed-rank tests and 10,000-sample bootstrap 95% CIs are computed on the 2,000-pair dumps of seed 42 (base vs. full on identical, exactly aligned pairs; each pair is the analysis unit). Table 12 reports the GVS contribution, and Table 13 the headline baseline comparisons.

Table 12: GVS contribution (2,000 pairs, seed 42): ADD(S)-0.1d without vs. with GVS, per-pair paired Wilcoxon p, and 10,000-sample bootstrap 95% CI of the gain. Top block: default Grounded-SAM localization; bottom block: detector-free alignment variant.

Dataset	w/o GVS	w/ GVS	Wilcoxon p	95% CI
Grounded-SAM localization				
YCBV	54.25	**58.50**	7.2×10^{-16}	[3.11, 5.43]
NOCS	82.65	**86.20**	1.8×10^{-7}	[2.01, 5.08]
TOYL	79.65	**83.20**	2.4×10^{-7}	[1.56, 5.59]
LM	52.80	**55.60**	1.6×10^{-7}	[1.59, 4.01]
Talk2DINO detector-free variant				
YCBV	30.15	**41.90**	2.4×10^{-44}	[9.93, 13.58]
NOCS	43.90	**61.25**	7.1×10^{-62}	[15.12, 19.53]
TOYL	58.90	**69.30**	1.0×10^{-26}	[7.71, 13.18]
LM	24.10	**35.00**	2.2×10^{-31}	[8.71, 13.00]

Table 13: Headline comparison per-pair statistics (2,000 pairs, seed 42, Grounded-SAM localization): ADD(S)-0.1d advantage of Ground2Pose over each baseline, paired Wilcoxon p, and 10,000-sample bootstrap 95% CI of the per-pair gain. Advantages are computed from unrounded values.

Baseline	Dataset	Advantage (pp)	Wilcoxon p	95% CI
vs. LatentFusion				
	YCBV	+34.2	$< 10^{-75}$	[31.5, 36.8]
	NOCS	+67.5	$< 10^{-75}$	[64.9, 70.0]
	TOYL	+28.6	$< 10^{-75}$	[24.4, 32.7]
	LM	+32.9	$< 10^{-75}$	[30.1, 35.6]
vs. GigaPose				
	YCBV	+32.5	$< 10^{-100}$	[29.5, 35.4]
	NOCS	+66.3	$< 10^{-100}$	[63.8, 68.9]
	TOYL	+47.0	$< 10^{-100}$	[43.1, 50.8]
	LM	+40.4	$< 10^{-100}$	[37.6, 43.2]

H Systematic Pilot Studies and Design Probes

To systematically investigate the combinatorial design space and validate our design principles without incurring prohibitive computational overhead, we conducted structured pilot sweeps across representative stratified subsets (96–192 evaluation pairs per benchmark). These exploratory probes serve to diagnose algorithmic alternatives, analyze sensitivity plateaus, and evaluate eight pre-registered negative-result hypotheses. The finalized configuration ($K_{\text{iter}} = 1, \gamma = 0.7, K_{\text{top}} = \text{all}$) is subsequently validated on the full 2,000-pair benchmarks with complete statistical significance testing.

H.1 Hyperparameter Sensitivity and Flatness (YCBV)

Table 14 explores hyperparameter sensitivity across trimming fraction γ, iteration count K_{iter}, and top-K correspondence retention on YCBV (defaults: $K_{\text{iter}} = 1$, $K_{\text{top}} = \text{all}$, $\gamma = 0.7$). The trimming ratio demonstrates a wide performance plateau across $\gamma \in [0.5, 0.9]$, establishing that GVS is robust to parameter selection and does not overfit to specific datasets.

H.2 Verification Iterations (K_{ITER})

Table 15 evaluates verification iterations $K_{\text{iter}} \in \{1, 2, 3\}$ on 192-pair subsets across all four benchmarks. A single iteration ($K_{\text{iter}} = 1$) captures the vast majority of achievable geometric improvement; additional iterations incur extra computational latency and risk over-trimming valid correspondences in dense regions. Hence, $K_{\text{iter}} = 1$ is adopted as the universal default.

Table 15: ADD(S)-0.1d across verification iterations K_{iter} under default Grounded-SAM configuration (192-pair diagnostic subsets per dataset).

K_{iter}	YCBV	NOCS	TOYL	LM
1 (default)	55.73	**81.77**	81.77	**56.25**
2	**57.81**	81.25	**83.33**	**56.25**
3	**57.81**	80.73	81.77	53.12

H.3 Rationale for the Default Configuration

The default configuration (K_{iter}=1, γ=0.7, K_{top}=all) was fixed once from the diagnostic sweeps above (Tables 14 and 15) and applied unchanged to all benchmarks, all localization variants, and all seeds; no parameter is re-tuned per dataset. *Keep ratio* $\gamma = 0.7$. ADD(S)-0.1d stays within a 1.1-point

Table 14: Hyperparameter sensitivity analysis on YCBV (96-pair diagnostic subsets). The top two blocks evaluate verification iterations (K_{iter}) and correspondence retention (K_{top}) on the detector-free Talk2DINO alignment variant (weak localization baseline); the bottom block evaluates the trimming keep ratio (γ) under default Grounded-SAM localization. Defaults: $K_{\text{iter}} = 1$, $K_{\text{top}} = \text{all}$, $\gamma = 0.7$.

Setting	ADD(S)-0.1d	Mask IoU
Diagnostic sweep on detector-free variant (weak localization):		
$K_{\text{iter}} = 0$	27.6	29.1
$K_{\text{iter}} = 1$ (default)	34.4	29.1
$K_{\text{iter}} = 2$	37.2	29.1
Top-K correspondence selection (detector-free variant):		
$K_{\text{top}} = \text{all}$ (default)	34.4	29.1
$K_{\text{top}} = 8$	33.5	29.1
$K_{\text{top}} = 16$	35.4	29.1
Trimming keep ratio γ under default Grounded-SAM:		
$\gamma = 0.5$	60.42	–
$\gamma = 0.6$	60.42	–
$\gamma = 0.7$ (default)	59.38	–
$\gamma = 0.8$	60.42	–
$\gamma = 0.9$	59.38	–
Adaptive (gap-based)	61.46	–

band across the full $\gamma \in [0.5, 0.9]$ sweep (Table 14, bottom block), i.e., the objective is flat over a wide band. We choose $\gamma = 0.7$ as a robust interior point of the plateau, maximizing the margin from both failure modes: too conservative ($\gamma \to 1$ keeps contaminated boundary matches) and too aggressive (small γ discards valid constraints). A gap-based adaptive γ improves YCBV but shows zero gain on LineMOD (negative probe 6 below), failing our cross-dataset adoption criterion. The plateau also transfers across datasets: under the full 2,000-pair protocol, moving γ to 0.5 or 0.8 changes ADD(S)-0.1d by at most 1.8 points (mean 0.8) on NOCS, TOYL, and LineMOD as well, confirming that no per-dataset re-tuning of γ is needed. *Iterations $K_{iter} = 1$.* A single trimming–re-solving pass captures the large majority of the geometric improvement (Table 15): the second iteration changes the result by at most $+2.1$ points (YCBV) while slightly regressing NOCS, and the third iteration degrades three of the four benchmarks through over-trimming. Since every iteration is an additional closed-form SVD, K_{iter}=1 is simultaneously the most economical choice. *Correspondence budget $K_{top} = all$.* Truncating the post-filter correspondence set to the top-K best-scored matches yields no consistent gain (34.4 / 33.5 / 35.4 for K_{top}=all / 8 / 16 on the diagnostic sweep) and discards matches that, although individually weaker, are required to constrain all six degrees of freedom of the pose—in particular rotation about the viewing axis. We therefore keep every correspondence that survives mask-geometric filtering and let the γ-quantile trimming, rather than a fixed budget, remove outliers.

H.4 Robust Registration Baselines (RANSAC-Kabsch, TEASER++, SC2-PCR)

To rigorously evaluate whether deterministic trimming outperforms stochastic and certifiable consensus under identical input evidence (Table 4), we implement matched robust-registration baselines. All baselines receive the identical 3D correspondence set generated by LoFTR after mask filtering (Eqs. equation 2–equation 3) and the identical outlier aperture: the geometric error threshold defined by the $\gamma = 0.7$ residual quantile of the initial full-set SVD fit. This isolates the solver itself as the only degree of freedom. The same configuration switches drive the two-factor experiment of Table 16: the mask factor is set to oracle, Grounded-SAM, or Talk2DINO correspondences, and the solver factor to plain Kabsch (verification off), RANSAC-Kabsch (256 iterations, shared aperture), or GVS; every cell is evaluated on the full 2,000 pairs (seed 42, pair-weighted).

Table 16: Two-factor evidence for the localization–geometry interface: mean ADD(S)-0.1d over the four benchmarks as a function of mask condition (rows) and solver (columns). The oracle row is a controlled intervention; segmenter rows are quasi-experiments at fixed geometry. The oracle row and the RANSAC column are new matched-protocol evaluations (2,000 pairs, seed 42); as a determinism check, the rerun GVS cells reproduce the per-benchmark values of Tables 4 and 3 to within 0.05 points on all four benchmarks, and the Kabsch and Grounded-SAM GVS cells are taken from those tables unchanged. Differences are computed from unrounded values.

Mask condition	Kabsch	RANSAC-256	GVS (Ours)	$\Delta_{\mathrm{GVS-K}}$
Oracle GT mask (intervention)	84.3	85.6	**87.7**	+3.3
Grounded-SAM (default; quasi-exp.)	67.3	68.6	**70.9**	+3.5
Talk2DINO (weak; quasi-exp.)	39.3	50.2	**51.9**	+12.6

RANSAC-Kabsch. At each iteration, a minimal subset of 3 point pairs is randomly sampled to compute a candidate rigid transformation $(\mathbf{R}, \mathbf{t})$ via Kabsch SVD. Inliers are scored against the shared aperture. After evaluating the iteration budget (32 or 256 iterations), the transformation is re-estimated via SVD on the maximal inlier consensus set, and all 2,000 pairs per benchmark are evaluated. RANSAC-Kabsch consistently trails deterministic GVS across all benchmarks because residual correspondence noise along object boundaries tends to be spatially coherent rather than uniformly distributed, making minimal 3-point sampling susceptible to perturbed boundary sets while deterministic full-set SVD provides a globally stable alignment anchor. *TEASER++ (certifiable).* TEASER++ Yang et al. (2021) replaces random sampling with truncated least squares over maximal cliques and graduated non-convexity (GNC), and certifies its solution against a bound on the outlier ratio. We run the official C++ implementation with scale estimation disabled (pure rigid motion), GNC-TLS rotation, and the noise bound set to the shared aperture, so that TEASER++ receives exactly the inlier definition used by GVS and RANSAC-Kabsch and the same assumed outlier budget $(1 - \gamma = 0.3)$. All 2,000 pairs per benchmark are evaluated. *SC2-PCR.* SC2-PCR Chen et al. (2022) expands seeds over a second-order spatial-compatibility measure and refits with weighted SVD. We use the official PyTorch implementation on the same filtered correspondence sets, with its spatial-compatibility threshold, seed-NMS radius, and refinement inlier threshold all re-expressed in the shared aperture (its native defaults are defined at scene scale), and all remaining hyperparameters at the official configuration (ρ=0.2, k_1=30, k_2=20, 10 power iterations). It costs 5.4 ms per solve at our correspondence scales ($\sim$34$\times$ GVS), and all 2,000 pairs per benchmark are evaluated. *Solver latency.* On the identical correspondence sets (13–305 points, median 52, RTX 3090): GVS solves in 0.16 ms per pair (deterministic full-set SVD plus one trimming pass); TEASER++ takes 0.45 ms at the median (up to 11 ms when GNC requires many iterations); SC2-PCR takes 5.4 ms; and our matched NumPy RANSAC-Kabsch takes 21.9 ms at 256 iterations. Deterministic trimming is thus not only variance-free but also the cheapest of the four solvers at object-scale correspondence densities. *Outcome.* As shown in Table 4, TEASER++ falls below even the untrimmed Kabsch baseline on every benchmark: its certificate bounds the outlier fraction but says nothing about how outliers are arranged, and it is the max-clique consensus it certifies—not the certificate itself—that presumes inliers dominate the correspondence graph; the outliers surviving mask filtering are spatially coherent boundary-band points that violate exactly this presumption and contaminate the initialization. The effect is structural rather than incidental—on YCBV, 99.4% of TEASER++'s successful pairs are also solved by RANSAC-Kabsch, while 28.7% of RANSAC-Kabsch successes fail under TEASER++. SC2-PCR's second-order spatial-consistency seeding is more tolerant of structured outliers: it outperforms TEASER++ on every benchmark, making it the strongest non-random-sampling baseline, yet it still trails both RANSAC-Kabsch and GVS everywhere—its consensus expansion remains anchored to the spectral seed basins selected at initialization—while costing $\sim$34$\times$ more per solve.

Table 17: Design-space comparison of robust registration solvers on our correspondence interface. RANSAC-Kabsch, TEASER++, and SC2-PCR are re-run under the matched protocol above; the trimmed-ICP row is positioned from its published formulation (nearest-neighbor correspondence re-assignment with point-to-plane trimming), as its iterate-to-fixed-point interface differs from our decoupled match-then-align pipeline. "Property" states the determinism/guarantee type; "Main limitation" is the dominant failure mode under our observed noise structure (coherent boundary-band outliers). Criterion entries are abbreviated: under the matched protocol each solver's threshold equals the shared γ=0.7 residual-quantile aperture.

Solver	Mechanism	Outlier criterion	Property	Main limitation
RANSAC-Kabsch	3-point sampling	γ-aperture	Probabilistic	Coherent outliers
TEASER++	Max-clique	Noise bound	Certifiable*	Contaminated clique
SC2-PCR	Spectral seeds	SC threshold	Seeded	Seed-basin overfit
Trimmed ICP	NN re-assignment	Point-to-plane trim	Local	Requires close init.
GVS (Ours)	Full-set SVD	3D residual trim	Deterministic	Outliers $> 1 - \gamma$

Operating envelope. GVS shares the assumption of γ-trimming estimators that the inlier fraction exceeds γ: when the outlier rate surpasses $1 - \gamma = 30\%$, the initial full-set fit can be dominated by outliers and the verifier inherits this bias. Within every regime measured in this paper, GVS is the most accurate and the cheapest solver (Tables 4, 16, and 17); beyond this envelope one should increase K_{iter} or consider certifiable estimators with a user-specified outlier bound (TEASER++), whose guarantee holds whenever that bound is respected—we do not measure alternatives beyond the envelope, so this suggestion rests on the certificate semantics rather than on our data. *Certifiable against an assumed bound on the outlier fraction and an explicit noise bound.

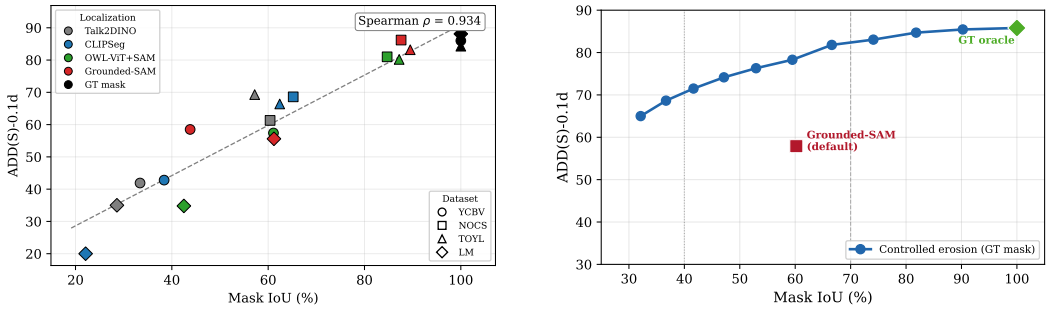
H.5 Analysis of Localization Failure Modes

A granular diagnostic probe (192 pairs per dataset, seed 42) indicates that localization failures in complex scenes are primarily dominated by *wrong-instance selection* (e.g., confusing distractor objects of similar categories) rather than subtle boundary erosion: mask IoU drops below 0.5 on 40.6% of YCBV and 32.8% of LineMOD pairs due to instance confusion. This is consistent with 2D semantic localization setting the empirical accuracy ceiling under controlled mask interventions, while GVS ensures that surviving correspondence inliers are optimally leveraged for 3D pose recovery.

H.6 Pre-Registered Negative-Result Probes ("Verification over Reweighting")

We systematically tested eight intuitive alternative designs under pre-registered adoption criteria ($\geq +1.0$ ADD gain on every evaluated benchmark without regression):

1. **Semantic correspondence reweighting.** Modulating LoFTR matches by DINOv2 patch similarity improves baseline Kabsch (+0.75 YCBV, +1.35 LM) but degrades the full GVS pipeline (-1.30 YCBV, -2.05 LM).
2. **Boundary-distance soft weighting.** Weighting matches by mask boundary distance loses 3–10 points under oracle mask erosion ($r \in \{0, 4, 8, 12\}$ px), as silhouette points carry critical rotational constraints.
3. **Pose-to-mask closed-loop feedback.** Re-prompting the segmenter using the predicted pose degrades ADD(S)-0.1d by -1.8 and -3.4 points on YCBV and LineMOD, respectively.
4. **Geometry-guided mask selection.** Replacing SAM's predicted-IoU mask ranking with downstream geometric inlier ranking yields identical performance on YCBV (58.0 vs. 58.0).
5. **Top-k instance rescue.** Geometrically verifying top-k candidate bounding boxes recovers only 11–19% of instance failures while degrading LineMOD by -4.7 points.
6. **Adaptive trimming fraction.** A gap-based adaptive γ improves YCBV (+2.1) but shows zero gain on LineMOD, failing our multi-dataset adoption threshold.
7. **SAM multimask arbitration.** Replacing SAM predicted-IoU candidate selection with a GVS solve-level residual score yields identical results on YCBV (+0.0) and regresses on LineMOD (-3.12).



(a) Mask quality predicts pose accuracy ($\rho = 0.934$). (b) Controlled erosion budget on YCBV.

Figure 4: **Controlled diagnosis of the localization–geometry interface.** (a) ADD(S)-0.1d versus mask IoU across 20 mask-source–dataset combinations (Table 3), exhibiting a strong rank correlation (Spearman $\rho = 0.934$), quantifying that localization quality sets the empirical accuracy ceiling. (b) Controlled boundary erosion of ground-truth masks on YCBV (100 pairs per erosion level): pose accuracy remains near the oracle level until IoU ≈ 0.7, then degrades rapidly. Actual segmenter markers sit below the boundary erosion curve because real localization errors are dominated by wrong-instance selections rather than pure boundary degradation.

8. **Attention-guided mask refinement.** Gating detector-free masks with DINOv2 self-attention yielded inconsistent dataset-dependent shifts (-2.3 YCBV to $+0.9$ NOCS) and was excluded.

Together, these findings establish a clear methodological principle: **robustness is most reliably achieved through hard 3D residual verification at the correspondence level, whereas heuristic soft reweighting or coarse-grained feedback introduces conflicting signals**.

I Localization–Geometry Interface Diagnosis

Fig. 4 provides a controlled diagnosis of the localization–geometry interface: (a) mask IoU versus pose accuracy across 20 mask-source–dataset combinations (Table 3), and (b) a controlled boundary-erosion budget on YCBV; details and interpretation are given in the caption.

J Qualitative Pose Estimation Results

Fig. 5 shows final pose estimation results across the four datasets, with predicted 3D bounding boxes rendered in blue against ground truth in green.

K Qualitative Mask Comparison across Segmenters

Fig. 6 provides qualitative visualizations comparing masks generated by the detector-free Talk2DINO alignment variant against our default Grounded-SAM configuration and ground truth across candidate pairs from YCB-Video and LineMOD. Grounded-SAM yields crisper and more often instance-pure masks even under heavy clutter, partial occlusions, and thin structures, explaining its superior downstream pose performance. In contrast, the detector-free variant occasionally bleeds onto adjacent distractors or background regions, supplying contaminated correspondences to the pose solver.

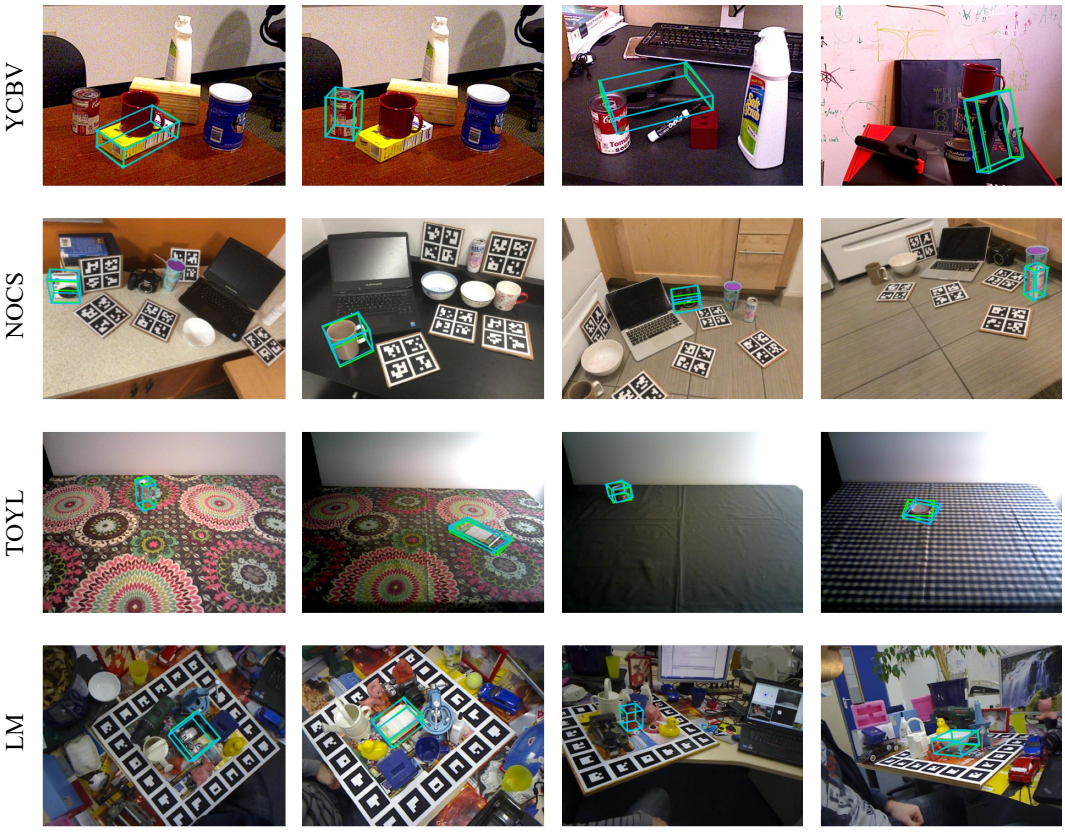


Figure 5: **Qualitative pose estimation results (our predictions).** Predicted 3D bounding boxes rendered in blue against ground truth in green. Rows: datasets (YCBV, NOCS, TOYL, LM); columns: four objects per dataset with successful predictions (per-case ADD well below $0.1d$), using our default localization.

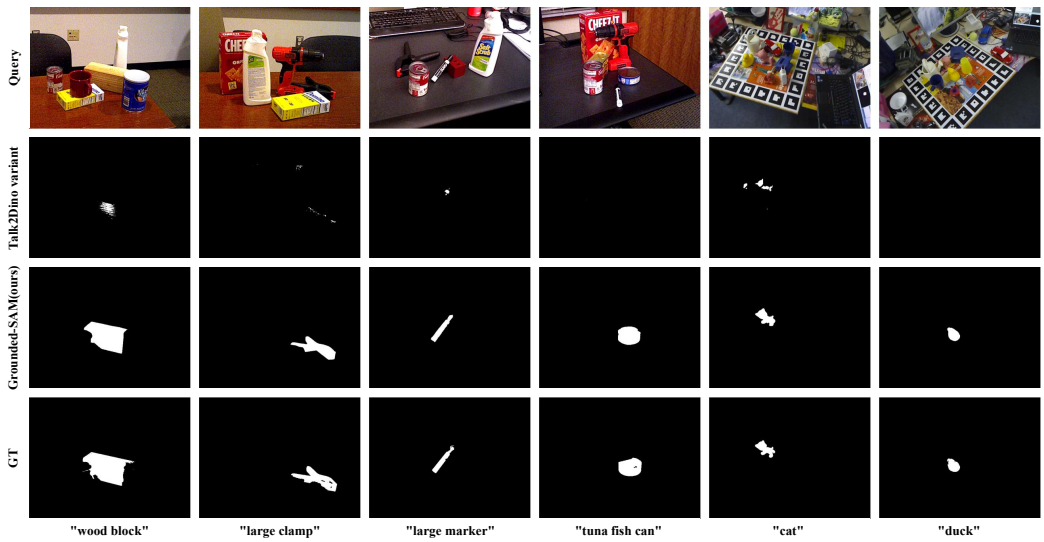


Figure 6: **Mask quality comparison across segmenters.** Rows: query RGB, detector-free alignment variant mask, our default Grounded-SAM mask, and ground truth. Columns: screened candidate pairs where the variant fails but Grounded-SAM succeeds, from YCBV (top four) and LineMOD (bottom two). Grounded-SAM masks closely match ground truth even on thin structures and cluttered backgrounds.