REVISITING CO-SALIENT OBJECT DETECTION VIA EXPLICIT SEMANTIC CONSENSUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Co-salient object detection (CoSOD) aims to locate and segment the common Co-salient object detection (CoSOD) aims to locate and segment the common salient object across a group of images. Existing methods infer group-level consensus in visual feature space and carry this implicit representation into object localization and segmentation. Yet visual similarity can cause semantically different distractors to enter the implicit consensus, reducing its reliability for segmentation. This motivates us to reconsider whether the essential shared information must remain encoded in a rich implicit visual representation, or can instead be preserved in a cleaner and more explicit form. We propose CoSEC (Co-Salient Explicit Consensus), a fully training-free framework that represents group-level consensus as a compact textual concept and uses it as the sole interface between group-level reasoning and per-image localization. A Semantic Consensus Estimator identifies the common object from the image group using a multi-image vision-language model, while a Consensus-Guided Segmenter independently grounds this concept in each image using a frozen text-conditioned segmenter. Without task-specific training or fine-tuning, CoSEC outperforms existing methods across all three benchmarks. On the challenging CoCA benchmark, it achieves an MAE of 0.023 and an $F_\beta^{\max}$ of 0.885, while requiring only 0.31 s/image on a single GPU. These results suggest that once the shared semantic identity is reliably established, a compact explicit semantic consensus can replace a rich implicit visual representation for strong CoSOD performance. Code will be released soon.

1 INTRODUCTION

Co-salient object detection (CoSOD) aims to segment the common salient objects shared across a group of images, with wide applications spanning image retrieval Papushoy & Borş (2015), weakly supervised segmentation Zeng et al. (2019), and visual data organization Cheng et al. (2014). Unlike single-image salient object detection (SOD) Borji et al. (2019), which defines saliency with respect to a single image's local context, CoSOD requires reasoning across the entire group to determine which objects are consistently and collectively prominent. This is inherently challenging: despite substantial variations in appearance across the group, the semantic identity of the shared object must be inferred collectively, while its spatial extent must be resolved in each individual image.

To address this challenge, existing CoSOD methods generally reason over visual features from the image group through different paradigms. Training-based approaches Fan et al. (2021); Jin et al. (2020); Yu et al. (2022); Zhang et al. (2020a) typically learn to identify recurring co-salient patterns by modeling relationships among visual features across the image group, while recent training-free methods Xiao et al. (2024); Yu et al. (2025); He et al. (2026) infer cross-image consistency from frozen visual features through heuristic feature-space operations. Despite their different mechanisms, both paradigms share a common representational pattern: the information used to identify the shared target remains encoded in visual features, forming an implicit consensus that guides subsequent segmentation.

However, as demonstrated in the left of Figure 1 (a), when target and distractor features overlap in the visual embedding space, the implicit consensus may include visually similar distractors while failing to cover semantically consistent but visually diverse target features. We further draw the consensus heatmap and corresponding predictions of previous implicit methods on the case of *Avocado*

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.

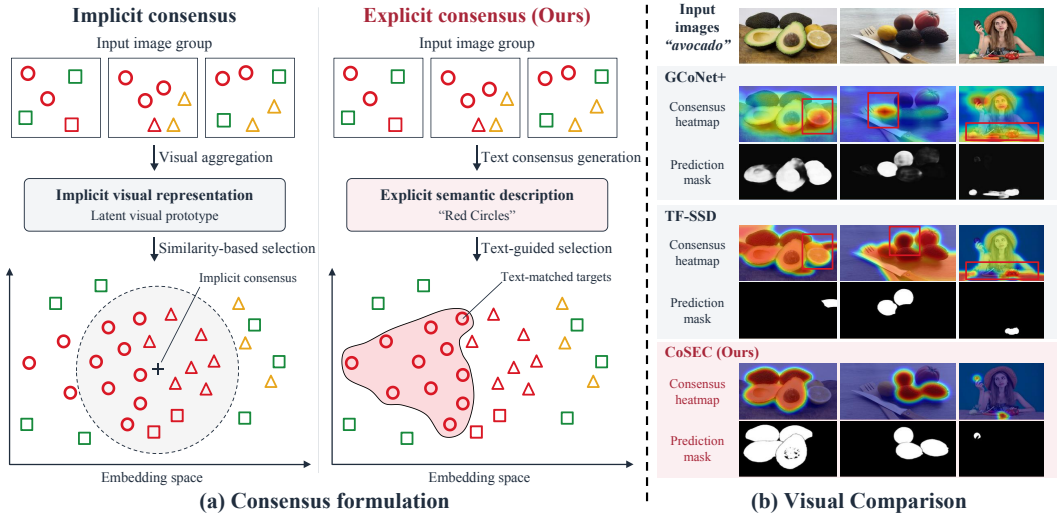


Figure 1: **Implicit visual consensus vs. explicit semantic consensus for CoSOD. (a) Schematic comparison of consensus formulations.**Circles represent shared targets; triangles and squares represent distractors. In implicit-consensus methods, similarity-based selection around a latent visual prototype can include distractors while missing some true targets. In contrast, CoSEC represents the shared target with an explicit semantic description (e.g., "Red Circles") that guides per-image localization. **(b) Visual Comparison.** Consensus heatmaps and predicted masks from the training-based GCoNet+ Zheng et al. (2023), training-free TF-SSD He et al. (2026), and CoSEC are shown for an image group sharing avocados. Red boxes highlight responses to distractors; CoSEC produces more focused target responses and more accurate predicted masks in this example.

in Figure 1 (b). The "red box" indicates the erroneous responses in the consensus heatmap. It can be found that previous methods assign strong responses to visually similar distractors, such as lemons, rather than consistently attending to the shared avocados in the heatmaps and cause incorrect predictions. These cases suggest a mismatch between visual-feature proximity and the agreement CoSOD requires. Instances of the shared object can look markedly different across images, while semantically unrelated distractors can resemble the target. Rather than attributing such failures solely to inaccurate consensus inference, we take this observation as motivation to reconsider the representation itself: *can group-level consensus explicitly state what is shared, without passing a visually entangled representation to the subsequent segmentation stage?*

Recent advances in vision-language models (VLMs) Bai et al. (2025); Li et al. (2023) and text-conditioned segmenters Kirillov et al. (2023); Carion et al. (2026) make such an explicit formulation practically realizable. VLMs can transform rich visual evidence into concise semantic consensus that explicitly identify the common object, while text-conditioned segmenters can directly ground a given concept in each image without requiring further cross-image interaction. Such explicit semantic consensus preserves *what* is shared across the group in a clear and compact form, without being confounded by visual similarities between targets and distractors, as illustrated in the right of Figure 1 (a). In this way, group-level reasoning and per-image spatial prediction are connected through a clean, explicit semantic interface rather than a complex implicit visual representation.

Building on this insight, we introduce **Co**-**S**alient **E**xplicit **C**onsensus (**CoSEC**), a fully training-free framework that realizes group-level consensus as a concise textual concept to explicitly guide co-salient object detection. Specifically, CoSEC first employs a **Semantic Consensus Estimator (SCE)** to transform the visual evidence across the image group into an explicit textual consensus that identifies the common object. Crucially, this explicit consensus serves as the only group-level information passed to the subsequent stage. Such a shared textual interface eliminates the need for further cross-image feature interaction during segmentation, thereby avoiding interference caused by feature variations of the same object across different images. Guided by this consensus, we propose a **Consensus-Guided Segmenter (CGS)** that independently generates candidate masks for each image and applies an adaptive threshold strategy to select and merge them into the final co-saliency

predictions. As shown in Figure 1 (b), CoSEC concentrates its responses on the shared avocados precisely and produces more complete target masks in the same avocado group, illustrating the potential benefit of explicitly identifying the target before per-image segmentation.

Experiments show that CoSEC achieves SOTA (State-Of-The-Art) results compared with previous supervised, unsupervised, and training-free baselines across widely used datasets. On the challenging CoCA dataset, CoSEC achieves an MAE of 0.023 and an $F_\beta^{\max}$ of 0.885, while requiring only 0.31 s/image on a single GPU, approximately $9.3\times$ faster than TF-SSD and $2.7\times$ faster than ZS-CoSOD (Table 5). Moreover, strong performance is retained when we try to generate the consensus through independent per-image VLM predictions, suggesting that the effectiveness stems from the explicit-consensus formulation rather than a particular multi-image VLM.

Our contributions are as follows:

- We revisit CoSOD from the perspective of exploring a compact and explicit textual semantic consensus instead of a complex and implicit consensus with visually similar distractors. The explicit consensus is used as the only guidance of co-saliency segmentation, thereby avoiding interference caused by feature variations across different images.
- We propose CoSEC, a fully training-free framework with two designed modules: a Semantic Consensus Estimator (SCE) that produces a textual group consensus, and a Consensus-Guided Segmenter (CGS) that conditions on this consensus with an adaptive threshold strategy to produce co-saliency masks for each image independently.
- CoSEC achieves SOTA results on widely used datasets, outperforming all compared supervised, unsupervised, and training-free baselines. On the challenging CoCA dataset, achieves a 19.9% improvement in $F_\beta^{\max}$ over the previous SOTA training-free method, while also running roughly $9\times$ faster at 0.31 s/image.

2 Related Work

Co-Salient Object Detection. CoSOD identifies shared objects despite appearance changes and salient distractors, making cross-image consensus central to the task. Methods span supervised Jin et al. (2020); Fan et al. (2021); Yu et al. (2022); Zhang et al. (2021); Zheng et al. (2023); Zhu et al. (2023); Li et al. (2024); Wang et al. (2025), unsupervised Wang et al. (2023); Chakraborty et al. (2024); Chakraborty & Samaras (2024); Liu et al. (2023), and training-free Xiao et al. (2024); Yu et al. (2025); He et al. (2026) approaches. ICNet Jin et al. (2020) extracts intra-saliency prototypes and inter-image correlations, while GCoNet and GCoNet+ Fan et al. (2021); Zheng et al. (2023) encourage intra-group compactness and inter-group separability. CADC Zhang et al. (2021) uses dynamic kernels; DCFM Yu et al. (2022) generates democratic prototypes; CoRP Zhu et al. (2023) purifies co-representations; and CONDA Li et al. (2024) learns condensed cross-image associations. VCP Wang et al. (2025) embeds visual consensus in learnable prompts through parameter-efficient tuning. Among unsupervised methods, TokenCut Wang et al. (2023), US-CoSOD Chakraborty et al. (2024), and SCoSPARC Chakraborty & Samaras (2024) avoid pixel annotations; the latter two use segmentation frequency statistics and patch- and region-level correspondences, respectively. Training-free ZS-CoSOD Xiao et al. (2024) generates group prompts for SAM Kirillov et al. (2023); PAP-SAM Yu et al. (2025) adds global-local priors; and TF-SSD He et al. (2026) filters SAM proposals using DINO Caron et al. (2021) attention across quality, intra-image, and inter-image stages. These methods typically infer shared targets through visual correspondences that may include visually similar distractors; CoSEC instead names the shared target before localization.

Vision-Language Models. Contrastive image–language pre-training Radford et al. (2021); Li et al. (2022b) links visual evidence to textual concepts and supports zero-shot transfer. Recent VLMs can reason over multiple images jointly and express shared content in natural language. Qwen2.5-VL Bai et al. (2025) and LLaVA-OneVision Li et al. (2025) support multi-image input, while BLIP-2 Li et al. (2023) provides per-image predictions that can be reconciled into group-level consensus. CoSEC uses Qwen2.5-VL by default and evaluates the alternatives to test dependence on the VLM and input setting. The resulting concise text serves as an interpretable interface to downstream segmentation.

Text-Guided Segmentation. Text-guided segmentation uses category names or natural-language descriptions to specify the regions to segment, providing a flexible semantic interface for dense prediction. A growing number of models explore this language-based formulation Wang et al. (2024); Jin et al. (2025; 2026); Zhu et al. (2026). LSeg Li et al. (2022a) aligns pixel features with text embeddings to generalize to unseen categories, while CLIPSeg Lüddecke & Ecker (2022) adapts CLIP with a dense prediction decoder to produce masks from text or image prompts. The recently introduced SAM 3 Carion et al. (2026) supports open-vocabulary concept-driven instance segmentation, producing scored masks for matching objects. CoSEC keeps SAM 3 frozen and connects group-level semantics to per-image segmentation without further cross-image feature aggregation.

3 METHOD

3.1 OVERVIEW

Conventional CoSOD methods aggregate visual features across images into an implicit group-level consensus for object localization and segmentation. Although this representation captures cross-image relationships, it may assign high similarity to visually similar but semantically different distractors. We therefore separate group-level target identification from per-image segmentation, using an explicit description of the shared object as the interface. Figure 2 illustrates the CoSEC pipeline.

Given a group of N images $\mathcal{G} = \{I_i\}_{i=1}^{N}$, CoSEC employs a **Semantic Consensus Estimator (SCE)** to transform group-level visual evidence into an explicit semantic consensus c^*. Unlike implicit feature-space consensus, this compact textual description identifies the shared object and serves as the only group-level information passed to localization.

The **Consensus-Guided Segmenter (CGS)** then independently segments the target in each image using a frozen text-conditioned segmenter. For each image I_i, CGS generates candidate masks conditioned on c^* and applies an adaptive threshold strategy to select an image-specific confidence threshold, merging the retained masks into the final co-saliency prediction M_i.

3.2 SEMANTIC CONSENSUS ESTIMATOR (SCE)

SCE identifies the shared object across the image group and expresses its semantic identity as a textual consensus. This explicit consensus guides subsequent segmentation without carrying forward a feature-space representation that may entangle target instances with visually similar distractors.

To generate the explicit semantic consensus for image group $\mathcal{G}$, we employ a frozen vision-language model (VLM) $\mathcal{V}$, formulated as:

$$c^* = \mathcal{V}(\mathcal{G}; p), \tag{1}$$

where p is a fixed task instruction. All images in $\mathcal{G}$ are provided jointly to $\mathcal{V}$, allowing the model to identify the common object from the group as a whole. The instruction p asks for a short object name to focus the response on the shared identity rather than the details of any individual image:

> p : *Look at all the images together. What is the common salient object that appears across all these images? Answer with only the object name (one or two words).*

Importantly, this consensus is inferred without access to any predefined category vocabulary, ground-truth annotations, or dataset-specific cues, making the entire process fully zero-shot.

3.3 CONSENSUS-GUIDED SEGMENTER (CGS)

Guided by the explicit consensus c^*, CGS localizes and segements the target co-salient object in each image independently. In detail, we use SAM 3 Carion et al. (2026) as a frozen text-conditioned segmenter $\mathcal{S}$, and query each image I_i with the same consensus label c^*. For each image, $\mathcal{S}$ produces a set of candidate co-saliency masks along with their confidence scores:

$$\{(M_i^{(j)},\, s_{ij})\}_{j=1}^{K_i} = \mathcal{S}(I_i,\, c^*), \tag{2}$$

where K_i is the number of candidates for image I_i, $M_i^{(j)}$ denotes the j-th candidate mask, and s_{ij} is its confidence score returned by SAM 3.

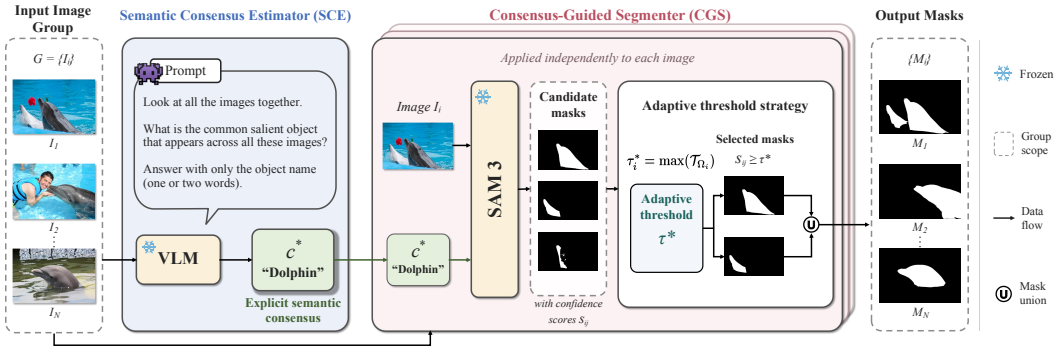


Figure 2: Overview of the CoSEC pipeline. The Semantic Consensus Estimator (SCE) uses a frozen VLM to identify the common object in the input image group and expresses it as a short text label ("Dolphin" in this example). This label is the only group-level information passed to the Consensus-Guided Segmenter (CGS). Guided by the label, frozen SAM 3 produces candidate masks and confidence scores for each image. CGS applies an image-specific adaptive threshold and merges the selected masks by pixel-wise union to produce the final co-saliency masks. The stacked CGS frames indicate that each image is processed independently.

Selecting only the highest-confidence mask can miss other instances or parts of the shared object. As shown in Figure 3 (a), the Top-1 mask covers only one beer bottle, while other bottles appear in separate candidate masks. Combining these complementary masks yields a more complete prediction. Yet a single cutoff is unsuitable across images: Figure 3 (b) shows that the per-image mean of the four highest candidate scores spans much of the confidence range across the 1,295 CoCA images. A fixed threshold may therefore discard valid masks in low-confidence images, whereas uniformly lowering it may admit unreliable proposals in other images.

To address this, we use an **adaptive threshold strategy** to select a confidence threshold for each image. Specifically, for the candidate masks of each image, let s_i^* denote the highest confidence score. We then define a candidate threshold set $\boldsymbol{\mathcal{T}} = \{\tau_k\}_{k=1}^{K}$, shared across all datasets, where K denotes the size of the threshold set, and form the image-specific subset containing all thresholds strictly below s_i^*:

$$\boldsymbol{\mathcal{T}}_{\Omega_i} = \left\{\tau_k \in \boldsymbol{\mathcal{T}} \mid \tau_k < s_i^*\right\}. \tag{3}$$

When this subset is non-empty, we choose its largest element as the threshold for image I_i:

$$\tau_i^* = \max\left(\boldsymbol{\mathcal{T}}_{\Omega_i}\right), \tag{4}$$

where $\max(\cdot)$ denotes the largest element of a set. This choice retains the strictest available threshold below the highest candidate score, adapting to the confidence level of each image without introducing learned parameters.

We then retain all candidate masks with scores at least τ_i^* and take their pixel-wise union to obtain the final prediction:

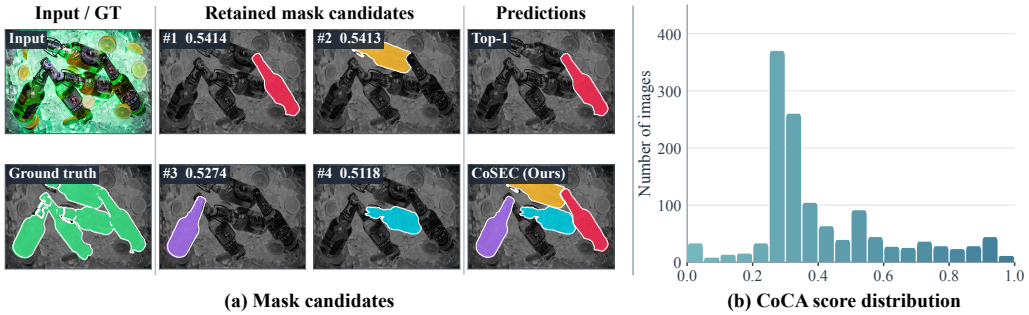
$$M_i = \bigcup_{j:\, s_{ij} \geq \tau_i^*} M_i^{(j)}. \tag{5}$$

Here, $\bigcup$ denotes the union of foreground pixels across the retained masks. If the segmenter returns no candidates or $\boldsymbol{\mathcal{T}}_{\Omega_i}$ is empty, we set $M_i = \mathbf{0}$. This strategy adapts to confidence variations across images while allowing multiple masks to contribute to a more complete co-saliency prediction. Applying CGS independently to all images in $\mathcal{G}$ yields the final set of co-saliency masks $\{M_i\}_{i=1}^{N}$ for the image group.

4 Experiments

4.1 Datasets and Metrics

We evaluate on the three standard CoSOD benchmarks: CoCA Zhang et al. (2020b) (the most challenging, containing extraneous salient distractors), CoSal2015 Zhang et al. (2015), and



(a) Mask candidates **(b) CoCA score distribution**

Figure 3: Mask coverage and confidence-score variation in the CoCA Beer Bottle group. **(a)** Input, ground truth, four retained mask candidates, and the Top-1 and CoSEC (Ours) predictions for one image in the group. Selecting only the Top-1 mask misses other bottle instances and covers 20% of the ground-truth target pixels, whereas our adaptive threshold strategy retains complementary candidates and raises coverage to 82%. **(b)** Empirical distribution of mean Top-4 SAM 3 confidence scores across CoCA. For each of the 1,295 images, we average the scores of its four highest-scoring mask candidates. The horizontal axis shows this per-image mean, and the vertical axis counts images in each score interval. The broad spread indicates substantial confidence variation across images.

CoSOD3k Fan et al. (2020). We report the four standard saliency metrics: mean absolute error (MAE, lower is better), maximum F-measure ($F_\beta^{\max}$) Fan et al. (2022), maximum E-measure ($E^{\max}$) Fan et al. (2018), and structure measure (S_α) Fan et al. (2017).

4.2 Implementation

We use Qwen2.5-VL-3B-Instruct Bai et al. (2025) as SCE's default multi-image VLM (CoSEC (Qwen)). All images in a group are provided jointly in a single query, allowing the VLM to identify their shared salient object. We also evaluate LLaVA-OneVision-7B Li et al. (2025) under the same multi-image setting to examine whether the results depend on the choice of VLM (CoSEC (LLaVA)). Furthermore, we explore whether a group-level semantic consensus can also be derived from single-image predictions. We evaluate BLIP-2 Li et al. (2023), LLaVA-OneVision-7B, and Qwen2.5-VL-3B-Instruct in this setting: each image is queried independently, and a group-level text consensus is selected from the resulting predictions. The single-image procedure and its implementation details are described in Appendix C. We name our CoSEC as CoSEC (BLIP-single), CoSEC (LLaVA-single), and CoSEC (Qwen-single), repectively. Across all variants, CGS uses SAM 3 Carion et al. (2026) as the open-vocabulary segmenter with the same adaptive threshold strategy. All models are used off-the-shelf with no fine-tuning.

4.3 Quantitative and Qualitative Comparison

Quantitative Comparison. We compare CoSEC with supervised DCFM Yu et al. (2022), CoRP Zhu et al. (2023), GCoNet+ Zheng et al. (2023), CONDA Li et al. (2024), and VCP Wang et al. (2025); unsupervised TokenCut Wang et al. (2023), DVFDVD Amir et al. (2022), US-CoSOD Chakraborty et al. (2024), and SCoSPARC Chakraborty & Samaras (2024); and training-free ZS-CoSOD Xiao et al. (2024) and TF-SSD He et al. (2026). As shown in Table 1, multi-image CoSEC (Qwen) leads all baselines on all four metrics across three benchmarks. On distractor-rich CoCA, its $F_\beta^{\max}$ reaches 0.885 versus 0.686 (TF-SSD), while MAE drops from 0.069 (VCP) to 0.023. Its $F_\beta^{\max}$ reaches 0.953 on CoSal2015 and 0.874 on CoSOD3k. With LLaVA as the multi-image VLM, CoSEC attains $F_\beta^{\max}$ scores of 0.825, 0.943, and 0.855 on CoCA, CoSal2015, and CoSOD3k, showing its applicability across VLM backbones.

Furthermore, we investigate whether the explicit consensus can be obtained without joint multi-image VLM input. We query each image independently with BLIP-2, LLaVA, or Qwen and select a group-level text consensus from their predictions. All three single-image variants exceed the strongest previous state-of-the-art $F_\beta^{\max}$ on each benchmark; on CoCA, CoSEC (BLIP-single), CoSEC (LLaVA-single), and CoSEC (Qwen-single) reach 0.882, 0.902, and 0.890, respectively. In

Method	Type	CoCA Zhang et al. (2020b)				CoSal2015 Zhang et al. (2015)				CoSOD3k Fan et al. (2020)			
		MAE↓	$F_\beta^{\max}$ ↑	$E^{\max}$ ↑	S_α ↑	MAE↓	$F_\beta^{\max}$ ↑	$E^{\max}$ ↑	S_α ↑	MAE↓	$F_\beta^{\max}$ ↑	$E^{\max}$ ↑	S_α ↑
DCFM$_{\text{CVPR 2022}}$	S	0.085	0.598	0.783	0.710	0.067	0.856	0.892	0.838	0.067	0.805	0.874	0.810
CoRP$_{\text{TPAMI 2023}}$	S	0.103	0.597	0.769	0.715	0.055	0.882	0.912	0.867	0.060	0.827	0.890	0.838
GCoNet+$_{\text{TPAMI 2023}}$	S	0.081	0.637	0.814	0.738	0.056	0.891	0.924	0.881	0.062	0.834	0.901	0.843
CONDA$_{\text{ECCV 2024}}$	S	0.089	0.685	0.839	0.763	0.045	0.908	0.944	0.900	0.056	0.853	0.911	0.862
VCP$_{\text{CVPR 2025}}$	S	0.069	0.680	0.829	0.774	0.037	0.920	0.944	0.911	0.049	0.868	0.868	0.874
TokenCut$_{\text{CVPR 2022}}$	U	0.167	0.467	0.704	0.627	0.139	0.805	0.857	0.793	0.151	0.720	0.811	0.744
DVFDVD$_{\text{ECCVW 2022}}$	U	0.223	0.422	0.592	0.592	0.092	0.777	0.842	0.809	0.104	0.722	0.819	0.773
US-CoSOD$_{\text{WACV 2024}}$	U	0.116	0.546	0.743	0.672	0.070	0.845	0.886	0.840	0.076	0.779	0.861	0.801
SCoSPARC$_{\text{ECCV 2024}}$	U	0.092	0.614	0.782	0.711	0.062	0.869	0.905	0.851	0.064	0.827	0.889	0.823
ZS-CoSOD$_{\text{ICASSP 2024}}$	TF	0.115	0.549	–	0.667	0.101	0.799	–	0.785	0.117	0.691	–	0.723
TF-SSD$_{\text{CVPR 2026}}$	TF	0.077	0.686	0.815	0.763	0.089	0.899	0.926	0.890	0.090	0.860	0.908	0.861
CoSEC (BLIP-single)	TF	0.030	0.882	0.933	0.899	0.026	0.940	0.948	0.921	0.054	**0.877**	**0.915**	0.878
CoSEC (LLaVA-single)	TF	**0.022**	**0.902**	**0.948**	**0.915**	0.024	0.937	0.945	0.918	0.050	0.871	0.909	0.875
CoSEC (Qwen-single)	TF	0.024	0.890	0.943	0.908	0.021	0.950	0.958	0.929	0.048	**0.877**	0.913	0.880
CoSEC (LLaVA)	TF	0.032	0.825	0.897	0.874	0.024	0.943	0.956	0.927	0.044	0.855	0.898	0.875
CoSEC (Qwen)	TF	0.023	0.885	0.940	0.909	**0.020**	**0.953**	**0.963**	**0.933**	**0.040**	0.874	0.912	**0.885**

Table 1: Quantitative comparison on three CoSOD benchmarks. S, U, and TF denote supervised, unsupervised, and training-free methods, respectively. For CoSEC, parentheses specify the VLM used for consensus estimation; "single" denotes per-image VLM queries followed by selection of a group-level text consensus. For LLaVA-single, k=6. Gray rows denote CoSEC variants. Best reported results are **bold**; displayed ties are both bold.

particular, LLaVA-single achieves the best CoCA results across all four metrics. These results show that the explicit-consensus remains effective when the VLM processes images independently.

Qualitative Comparison. Figure 4 compares CoSEC with training-based GCoNet+, SCoSPARC, and VCP, as well as prior training-free TF-SSD and ZS-CoSOD, on four CoCA categories. CoSEC detects the small baseballs and moons and preserves the thin belt and fishing-rod structures. In particular, TF-SSD and VCP often segment the visually prominent people in "Baseball" and "Fishing Rod", whereas CoSEC isolates the shared targets; ZS-CoSOD largely misses them. These examples illustrate the benefit of explicitly identifying the shared target before segmentation.

4.4 Ablation Study

Adaptive Threshold Strategy in CGS. To validate the effectiveness of our adaptive threshold strategy in CGS, we use our default CoSEC (Qwen) configuration and vary only how CGS constructs the final mask: retaining the highest-scoring candidate, unioning candidates above a fixed threshold, or selecting an image-specific threshold from $\boldsymbol{\mathcal{T}}$ before union. Table 2 shows that adding lower fallback values to $\boldsymbol{\mathcal{T}}$ improves $F_\beta^{\max}$ on all three benchmarks. The full set $\boldsymbol{\mathcal{T}}=\{0.2, 0.3, 0.4, 0.5\}$ performs best among the adaptive configurations and exceeds every tested fixed threshold. On CoSOD3k, it gains 0.035 over fixed τ=0.5 and narrowly exceeds the highest-score-only baseline (0.874 vs. 0.873). Relative to fixed τ=0.5, it reduces empty-mask rates from 5.6% to 3.3% on CoCA, 2.9% to 1.1% on CoSal2015, and 10.9% to 7.0% on CoSOD3k. These results support image-specific threshold selection with multiple fallback levels for mask assembly.

Effect of Number of Input Images for SCE. To test how much group context SCE needs, we vary N, the number of images jointly supplied to SCE, while keeping Qwen, its instruction, and the SAM 3-based CGS unchanged. For $N \in \{2, 3, 4, 5\}$, we randomly select N images from each CoCA group for the SCE query; CGS then segments every image in the original group, and evaluation covers the full group. The "All" setting supplies the entire group to SCE. Figure 5 shows that $F_\beta^{\max}$ rises from 0.690 at N=2 to 0.879 at N=5, then reaches 0.885 with all images. This trend is consistent with more group views supporting more reliable semantic consensus, while the N=2 result shows that the pipeline can operate with limited group context.

Sensitivity to Prompt Design. We also evaluate different instruction designs with our CoSEC (Qwen). We propose five different instructions as: (i) **Paraphrase:** rewording the default instruction p in section 3.2; (ii) **Distractor-aware:** explicitly ignoring visually similar distractors; (iii) **Chain-of-thought:** ask the VLM to describe the target first then output c^* based on the description; (iv) **No-constraint:** removing the answer-length requirement; and (v) **Single-keyword:** requiring the output be one word. The complete prompts are in Appendix A.2. As shown in Table 3, the baseline

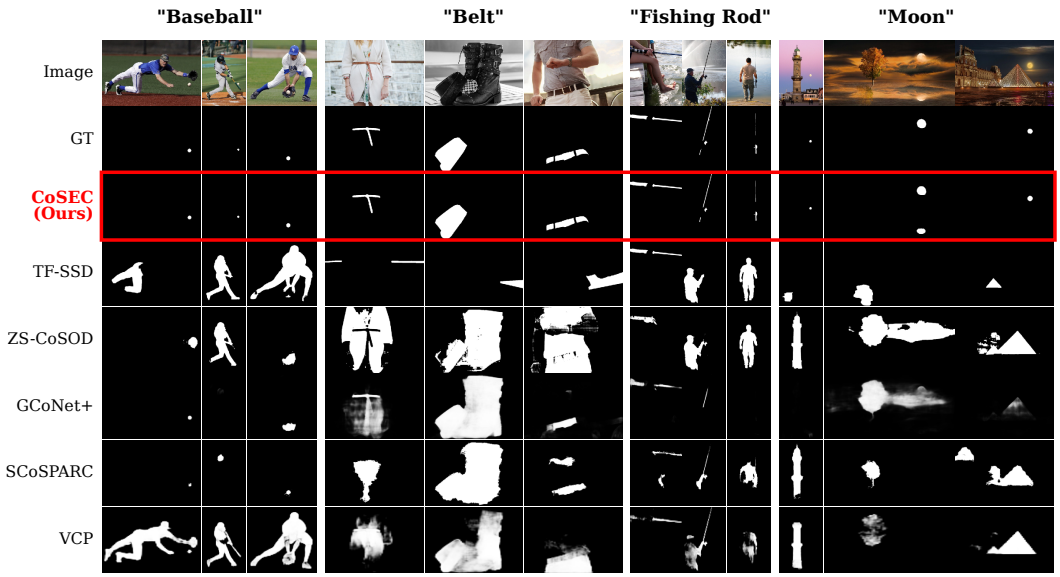


Figure 4: Qualitative comparison on four CoCA categories. **CoSEC (Ours)** is highlighted in red. Compared methods: TF-SSD He et al. (2026), ZS-CoSOD Xiao et al. (2024), GCoNet+ Zheng et al. (2023), SCoSPARC Chakraborty & Samaras (2024), VCP Wang et al. (2025).

Configuration	CoCA	CoSal2015	CoSOD3k
Highest-score only	0.848	0.942	0.873
Fixed τ=0.2	0.856	0.940	0.858
Fixed τ=0.3	0.871	0.943	0.853
Fixed τ=0.5	0.868	0.936	0.839
$\mathcal{T}$=$\{0.4, 0.5\}$	0.877	0.942	0.850
$\mathcal{T}$=$\{0.3, 0.4, 0.5\}$	0.884	0.949	0.861
$\mathcal{T}$=$\{0.2, 0.3, 0.4, 0.5\}$	**0.885**	**0.953**	**0.874**

Table 2: Ablation of the adaptive threshold strategy in CGS ($F_\beta^{\max}$ ↑). Full metrics are reported in Appendix A.1.

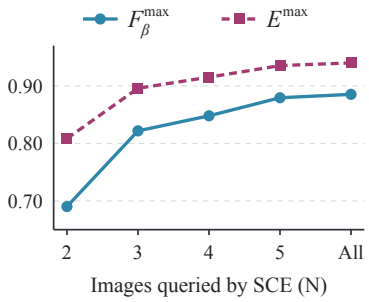


Figure 5: CoCA performance vs. k images randomly supplied to SCE.

achieves the highest accuracy on CoSOD3k and ties with the single-keyword variant on CoSal2015, while paraphrasing improves CoCA accuracy from 82.5% to 85.0%. These results support the use of a concise, direct object-naming instruction for semantic consensus estimation.

Inference Efficiency. Tables 5 and 6 report the inference efficiency of CoSEC on CoCA and its module-wise time breakdown, measured on a single NVIDIA RTX 4090 D GPU. Our default CoSEC (Qwen) averages 0.31 s per image for end-to-end inference, corresponding to speedups of 2.7× over ZS-CoSOD and 9.3× over TF-SSD while achieving the highest $F_\beta^{\max}$ among the compared methods. Of the 4.94 s average group processing time, SCE takes 0.57 s for a single joint VLM query, while CGS takes 4.37 s with images processed sequentially. With an average of 16.2 images per group, the SCE cost is amortized to approximately 0.04 s per image, compared with 0.27 s for CGS. Thus, group-level semantic consensus estimation contributes a small fraction of the total runtime, with per-image localization accounting for most of the cost.

4.5 Oracle Analysis

To examine the impact of consensus quality, we evaluate an oracle that replaces the VLM prediction with the ground-truth category name while keeping CGS fixed. Table 4 shows that this oracle raises $F_\beta^{\max}$ from 0.885 to 0.928 on CoCA and from 0.874 to 0.901 on CoSOD3k, demonstrating the

Prompt variant	CoCA	CoSal2015	CoSOD3k
Baseline	82.5	**76.0**	**80.0**
Paraphrase	**85.0**	72.0	75.0
Distractor-aware	82.5	74.0	74.4
Chain-of-thought	72.5	68.0	63.1
No-constraint	82.5	72.0	63.7
Single-keyword	80.0	**76.0**	78.1

Table 3: SCE prompt ablation (consensus accuracy in %; best values in bold).

Dataset	Metric	Single	CoSEC	Oracle
CoCA	$F_\beta^{\max}$	0.890	0.885	**0.928**
	MAE	0.024	0.023	**0.018**
CoSal2015	$F_\beta^{\max}$	0.950	**0.953**	0.898
	MAE	0.021	**0.020**	0.034
CoSOD3k	$F_\beta^{\max}$	0.877	0.874	**0.901**
	MAE	0.048	0.040	**0.033**

Table 4: Consensus-source ablation with fixed CGS. Oracle: ground-truth category name.

Method	Time (s/img)↓	$F_\beta^{\max}$ ↑
ZS-CoSOD Xiao et al. (2024)	0.83	0.549
TF-SSD He et al. (2026)	2.87	0.686
CoSEC (Ours)	**0.31**	**0.885**

Table 5: Inference time vs. CoCA $F_\beta^{\max}$ among training-free methods .

Stage	Per-group (s)	Per-image (s)
SCE (VLM)	0.57	0.04
CGS (SAM3)	4.37	0.27
Total	**4.94**	**0.31**

Table 6: Per-stage time breakdown of CoSEC on CoCA benchmark.

influence of VLM-predicted consensus on segmentation. Interestingly, on CoSal2015, CoSEC exceeds the category-name oracle, reaching $F_\beta^{\max}$ of 0.953 versus 0.898 and lower MAE (0.020 versus 0.034). One possible explanation is that some dataset category names are overly specific or specialized, while the VLM's broader descriptions may better align with SAM 3's learned text–visual associations. Effective consensus therefore requires not only correct identification of the shared object, but also a textual expression that the segmenter can reliably associate with the corresponding image regions. The category-name oracle is thus not a strict upper bound: VLM-derived consensus can guide segmentation more effectively than canonical labels. This suggests that semantic alignment across different foundation models could be a promising direction for future research.

4.6 Limitations

CoSEC's performance is inherently bounded by c^* quality: an incorrect VLM prediction propagates to all group images without self-correction. Its performance is also constrained by SAM 3's localization and segmentation capabilities, which can lead to empty masks for small or atypically textured objects even with a correct c^*. Despite these limitations, existing foundation models encode rich semantic knowledge, and We hope CoSEC will inspire future work to further leverage this rich semantic knowledge for CoSOD and similar visual understanding tasks.

5 Conclusion

We presented CoSEC, a training-free CoSOD framework that represents group-level consensus as an explicit, human-readable textual concept c^*, decoupling semantic consensus estimation from per-image segmentation. Built from frozen vision-language and segmentation foundation models, CoSEC requires no CoSOD-specific training, pixel-level annotations, or fine-tuning. It outperforms the compared supervised, unsupervised, and training-free baselines on CoCA, CoSal2015, and CoSOD3k while maintaining fast inference. Experiments with different VLMs and both joint-image and independent single-image inputs further support the effectiveness of the explicit-consensus framework. Oracle analysis highlights the importance of consensus accuracy and its compatibility with the segmenter, alongside the segmentation model's own limitations. Together, these findings show that a compact textual consensus can effectively convey shared target identity to guide independent per-image segmentation. By rethinking consensus representation in CoSOD, our work highlights the potential of explicit semantic guidance and we hope our CoSEC can encourage broader exploitation of the rich semantic knowledge already encoded in VLMs for CoSOD and related tasks.

AI USE STATEMENT

Generative AI tools were used solely to refine the English expression and improve the clarity of this manuscript. All AI-assisted modifications were carefully reviewed and approved by the authors.

REPRODUCIBILITY STATEMENT

The main text and appendix document the experimental setup, evaluation protocol, and implementation details needed to reproduce our results. We also provide the core implementation in the supplementary material, with code comments identifying the key configurations and processing steps.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. In *ECCVW*, 2022. URL https://arxiv.org/abs/2112.05814.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Comput. Vis. Media*, 5(2):117–150, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *ICLR*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Souradeep Chakraborty and Dimitris Samaras. Self-supervised co-salient object detection via feature correspondences at multiple scales. In *ECCV*, pp. 231–250, 2024.

Souradeep Chakraborty, Shujon Naha, Muhammet Bastan, Amit Kumar K. C., and Dimitris Samaras. Unsupervised and semi-supervised co-salient object detection via segmentation frequency statistics. In *WACV*, pp. 332–342, January 2024.

Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. SalientShape: group saliency in image collections. *Vis. Comput.*, 30(4):443–453, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-Measure: A new way to evaluate foreground maps. In *ICCV*, pp. 4548–4557, 2017.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pp. 698–704, 2018.

Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *CVPR*, pp. 2919–2929, 2020.

Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4339–4354, 2022.

Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *CVPR*, pp. 12288–12298, 2021.

Zhijin He, Shuo Jin, Siyue Yu, Shuwei Wu, Bingfeng Zhang, Li Yu, and Jimin Xiao. TF-SSD: A strong pipeline via synergic mask filter for training-free co-salient object detection. In *CVPR*, pp. 32216–32225, 2026.

Shuo Jin, Siyue Yu, Bingfeng Zhang, Mingjie Sun, Yi Dong, and Jimin Xiao. Feature purification matters: Suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In *ICCV*, pp. 20291–20300, 2025.

Shuo Jin, Siyue Yu, Bingfeng Zhang, Chao Yao, Meiqin Liu, and Jimin Xiao. TALENT: Target-aware efficient tuning for referring image segmentation. In *CVPR Findings*, pp. 7472–7482, 2026.

Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, and Wei Guo. ICNet: Intra-saliency correlation network for co-saliency detection. In *NeurIPS*, volume 33, pp. 18749–18759, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900, 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742, 2023.

Long Li, Nian Liu, Dingwen Zhang, Zhongyu Li, Salman Khan, Rao Anwer, Hisham Cholakkal, Junwei Han, and Fahad Shahbaz Khan. CONDA: Condensed deep association learning for co-salient object detection. In *ECCV*, pp. 287–303, 2024.

Yan Liu, Tengpeng Li, Yang Wu, Huihui Song, and Kaihua Zhang. Self-supervised image co-saliency detection. *Comput. Electr. Eng.*, 105:108533, 2023.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, pp. 7086–7096, 2022.

Alex Papushoy and Adrian G Borş. Image retrieval based on query by saliency content. *Digit. Signal Process.*, 36:156–173, 2015.

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

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *ECCV*, pp. 315–332, 2024.

Jie Wang, Nana Yu, Zihao Zhang, and Yahong Han. Visual consensus prompting for co-salient object detection. In *CVPR*, pp. 9591–9600, 2025.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. TokenCut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15790–15801, 2023.

Haoke Xiao, Lv Tang, Bo Li, Zhiming Luo, and Shaozi Li. Zero-shot co-salient object detection framework. In *ICASSP*, pp. 4010–4014, 2024.

Jizhe Yu, Xiya Bu, Yu Liu, and Kaiping Xu. PAP-SAM: Global-local prior adaptive perception SAM for co-salient object detection. In *ICMR*, pp. 1720–1729, 2025.

Siyue Yu, Jimin Xiao, Bingfeng Zhang, and Eng Gee Lim. Democracy does matter: Comprehensive feature mining for co-salient object detection. In *CVPR*, pp. 979–988, 2022.

Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, pp. 7223–7233, 2019.

Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *CVPR*, pp. 2994–3002, 2015.

Ni Zhang, Junwei Han, Nian Liu, and Ling Shao. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection. In *ICCV*, pp. 4167–4176, 2021.

Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, and Yao Zhao. CoADNet: Collaborative aggregation-and-distribution networks for co-salient object detection. In *NeurIPS*, volume 33, pp. 6959–6970, 2020a.

Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *ECCV*, pp. 455–472, 2020b.

Peng Zheng, Huazhu Fu, Deng-Ping Fan, Qi Fan, Jie Qin, Yu-Wing Tai, Chi-Keung Tang, and Luc Van Gool. GCoNet+: A stronger group collaborative co-salient object detector. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10929–10946, 2023.

Hao Zhu, Shuo Jin, Wenbin Liao, Jiayu Xiao, Yan Zhu, Siyue Yu, and Feng Dai. VIP: Visual-guided prompt evolution for efficient dense vision-language inference. In *ICML*, 2026. URL https://openreview.net/forum?id=G1BYq33C3G.

Ziyue Zhu, Zhao Zhang, Zheng Lin, Xing Sun, and Ming-Ming Cheng. Co-salient object detection with co-representation purification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8193–8205, 2023.

APPENDIX

Overview. This appendix supplements the main paper with detailed ablation settings and results, class-wise semantic consensus evaluation, additional visualizations, and an analysis of consensus estimation from single-image inputs.

Contents *Page*

A **Details of Ablation Studies** . . . 13
A.1 Adaptive Threshold Strategy . . . 13
A.2 Complete Prompts for SCE . . . 13
A.3 Semantic Consensus Accuracy . . . 14
B **Additional Qualitative Results** . . . 20
C **Single-Image Consensus Estimation** . . . 21
C.1 Method . . . 22
C.2 Ablation on the Number of Candidate Prompts . . . 23

A DETAILS OF ABLATION STUDIES

A.1 ADAPTIVE THRESHOLD STRATEGY

	CoCA			CoSal2015			CoSOD3k		
Configuration	MAE↓	$F_\beta^{\max}$ ↑	S_α ↑	MAE↓	$F_\beta^{\max}$ ↑	S_α ↑	MAE↓	$F_\beta^{\max}$ ↑	S_α ↑
Highest-score only	0.036	0.848	0.854	0.030	0.942	0.918	0.063	0.873	0.848
Fixed τ=0.2	0.028	0.856	0.894	0.024	0.940	0.925	0.045	0.858	0.876
Fixed τ=0.3	0.025	0.871	0.901	0.023	0.943	0.928	0.045	0.853	0.874
Fixed τ=0.5	0.024	0.868	0.900	0.024	0.936	0.924	0.047	0.839	0.867
$\mathcal{T}$=$\{0.4, 0.5\}$	0.023	0.877	0.904	0.023	0.942	0.927	0.045	0.850	0.873
$\mathcal{T}$=$\{0.3, 0.4, 0.5\}$	0.023	0.884	0.908	0.021	0.949	0.931	0.043	0.861	0.878
$\mathcal{T}$=$\{0.2, 0.3, 0.4, 0.5\}$ **(ours)**	**0.023**	**0.885**	**0.909**	**0.020**	**0.953**	**0.933**	**0.040**	**0.874**	**0.885**

Table 7: Full results of the adaptive threshold strategy ablation across three benchmarks. MAE↓, $F_\beta^{\max}$ ↑, and S_α ↑.

Table 7 extends Table 2 with MAE and S_α, showing that adaptive threshold selection improves both pixel-level accuracy and structural quality. Compared with retaining only the highest-scoring mask, the full adaptive strategy reduces MAE and increases S_α on all three benchmarks. On CoSOD3k, for example, MAE decreases from 0.063 to 0.040 and S_α rises from 0.848 to 0.885, despite only a small gain in $F_\beta^{\max}$. Fixed thresholds also improve MAE and S_α over the highest-score-only baseline, but their relative performance varies across datasets and metrics. Adding lower fallback thresholds to $\mathcal{T}$ consistently improves or maintains all three metrics, with the full set achieving the best results, including ties. These results support adapting candidate selection to each image when assembling the final mask.

A.2 COMPLETE PROMPTS FOR SCE

The following six prompts are used in the prompt-design ablation in Table 3. All variants use the same joint-image input and greedy decoding, differing only in the instruction text.

`Baseline:` Look at all the images together. What is the common salient object that appears across all these images? Answer with only the object name (one or two words).

`Paraphrase:` Examine these images as a group. Name the one object type that shows up as the shared, most noticeable subject in every image. Reply with just the object's name (one to two words).

`Distractor-aware:` These images share one common salient object, but each image may also contain other secondary objects that are not shared across the group. Identify the single object that is salient and common to ALL images, ignoring anything that appears in only some of them. Answer with only the object name (one or two words).

`Chain-of-thought:` Look at all the images together. First, briefly note the salient object in each image, one short phrase per image. Then, on a new final line starting exactly with `Answer:`, state the single object name shared across all images (one or two words).

`No-constraint:` Look at all the images together. What is the common salient object that appears across all these images? Describe it briefly.

`Single-keyword:` Look at all the images together. What is the single common salient object shared across all these images? Answer in one word only.

A.3 Semantic Consensus Accuracy

We evaluate whether the consensus c^* predicted by the default Qwen2.5-VL-3B-Instruct SCE correctly identifies the shared target. The evaluation covers all 80, 50, and 160 classes in CoCA, CoSal2015, and CoSOD3k, respectively, with one prediction per class. Table 8 summarizes the results. Tables 9–11 provide every class name, predicted c^*, automatic matching result, mismatch type, review decision, and explanation.

Evaluation criteria. *Strict* uses automatic exact/substring matching. *+Synonyms* additionally accepts annotated alternative names and spelling or formatting variants, such as *UAV→drone* and *bow_tie→bow tie*, including some differences in naming specificity. *Reviewed* further accepts coarse or related descriptions on a case-by-case basis, such as *pigeon→bird*. All criteria use the same predictions and retain every class in the denominator. Reviewed accuracy thus includes accepted partial descriptions and does not measure exact category-name recovery.

Dataset	Classes	Strict	+Synonyms	Reviewed	Errors
CoCA	80	66 (82.5%)	70 (87.5%)	74 (92.5%)	6 (7.5%)
CoSal2015	50	38 (76.0%)	48 (96.0%)	48 (96.0%)	2 (4.0%)
CoSOD3k	160	128 (80.0%)	146 (91.2%)	154 (96.2%)	6 (3.8%)

Table 8: Class-wise semantic consensus accuracy. Entries show counts and percentages; Errors reports predictions rejected after manual review.

Class-wise annotations. In Tables 9–11, *Pass*/*Fail* denote automatic matching results. For unmatched predictions, *Accept* denotes an alternative accepted under +Synonyms, *Partial* a coarse or related description accepted only under Reviewed, and *Error* a rejected prediction. Type and Explanation describe each mismatch; *Variant*, *Format*, and *Near-syn.* denote spelling/word-separation variants, underscore/space differences, and near-synonyms, respectively. A dash indicates no additional annotation. CoSal2015 counts both *deer→antelope* and *viola→cello* as errors, retaining 96.0% reviewed accuracy.

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
1	Accordion	accordions	Pass	–	–	–
2	UAV	drone	Fail	Synonym	Accept	Drone denotes a UAV.
3	Yellow duck	duck	Pass	–	–	–
4	alarm clock	clock	Pass	–	–	–
5	avocado	avocado	Pass	–	–	–
6	backpack	backpack	Pass	–	–	–
7	baseball	baseball	Pass	–	–	–
8	beer bottle	beer	Pass	–	–	–
9	belt	belt	Pass	–	–	–

Continued on the next page.

Table 9 (continued): CoCA

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
10	binoculars	binoculars	Pass	–	–	–
11	boots	boots	Pass	–	–	–
12	butterfly	butterfly	Pass	–	–	–
13	calculator	calculator	Pass	–	–	–
14	camel	camel	Pass	–	–	–
15	camera	camera	Pass	–	–	–
16	candle	candle	Pass	–	–	–
17	chopsticks	chopsticks	Pass	–	–	–
18	clover	clover	Pass	–	–	–
19	dice	dice	Pass	–	–	–
20	dolphin	dolphin	Pass	–	–	–
21	doughnut	donut	Fail	Variant	Accept	Accepted as equivalent.
22	dumbbell	dumbbell	Pass	–	–	–
23	eggplant	vegetables	Fail	General	Error	Expected: eggplant.
24	faucet	water	Fail	Shift	Error	Names the outflow instead of the object.
25	fishing rod	fishing rod	Pass	–	–	–
26	frisbee	frisbee	Pass	–	–	–
27	gift box	gift	Pass	–	–	–
28	glasses	glasses	Pass	–	–	–
29	globe	globe	Pass	–	–	–
30	glove	glove	Pass	–	–	–
31	guitar	guitar	Pass	–	–	–
32	hammer	hammer	Pass	–	–	–
33	hammock	hammock	Pass	–	–	–
34	handbag	purse	Fail	Synonym	Accept	Accepted US English equivalent.
35	harp	harp	Pass	–	–	–
36	hat	hat	Pass	–	–	–
37	headphone	headphones	Pass	–	–	–
38	helicopter	helicopter	Pass	–	–	–
39	high heels	shoe	Fail	Hypernym	Partial	Related, but too broad.
40	hourglass	hourglass	Pass	–	–	–
41	ice cream	ice cream	Pass	–	–	–
42	key	key	Pass	–	–	–
43	lollipop	lollipop	Pass	–	–	–
44	macaroon	macaron	Fail	Variant	Accept	Accepted as equivalent.
45	microphone	microphone	Pass	–	–	–
46	minions	banana	Fail	Wrong	Error	Confuses the cartoon character with a banana.
47	moon	moon	Pass	–	–	–
48	persimmon	fruit	Fail	General	Error	Expected: persimmon.
49	pigeon	bird	Fail	Hypernym	Partial	Related, but insufficiently precise.
50	pillow	pillow	Pass	–	–	–
51	pine cone	pine cone	Pass	–	–	–
52	pineapple	pineapple	Pass	–	–	–
53	pocket watch	clock	Fail	Hypernym	Partial	Related, but insufficiently precise.
54	poker	playing card	Fail	Hypernym	Partial	Related, but insufficiently precise.
55	potato	potato	Pass	–	–	–
56	pumpkin	pumpkin	Pass	–	–	–
57	rabbit	rabbit	Pass	–	–	–
58	rocking horse	horse	Pass	–	–	–

Continued on the next page.

Table 9 (continued): CoCA

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
59	roller-skating	skateboard	Fail	Wrong	Error	Names equipment instead of the activity.
60	rolling pin	rolling pin	Pass	–	–	–
61	soap bubble	bubble	Pass	–	–	–
62	squirrel	squirrel	Pass	–	–	–
63	stethoscope	stethoscope	Pass	–	–	–
64	sticky note	pen	Fail	Wrong	Error	Confused by the scene context.
65	stool	stool	Pass	–	–	–
66	strawberry	strawberry	Pass	–	–	–
67	sunflower	sunflower	Pass	–	–	–
68	tablet	tablet	Pass	–	–	–
69	teddy bear	teddy bear	Pass	–	–	–
70	thermometer	thermometer	Pass	–	–	–
71	tomato	tomato	Pass	–	–	–
72	towel	towel	Pass	–	–	–
73	toy car	car	Pass	–	–	–
74	typewriter	typewriter	Pass	–	–	–
75	violin	violin	Pass	–	–	–
76	waffles	waffle	Pass	–	–	–
77	watering can	watering can	Pass	–	–	–
78	watermelon	watermelon	Pass	–	–	–
79	wheelchair	wheelchair	Pass	–	–	–
80	whisk	whisk	Pass	–	–	–

Table 9: Complete class-wise consensus predictions and review annotations on CoCA (80 classes). Type and Review follow the definitions in Section A.3.

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
1	aeroplane	airplane	Fail	Synonym	Accept	Accepted as equivalent.
2	apple	apple	Pass	–	–	–
3	axe	axe	Pass	–	–	–
4	babycrib	crib	Pass	–	–	–
5	banana	banana	Pass	–	–	–
6	baseball	baseball	Pass	–	–	–
7	bear	bear	Pass	–	–	–
8	billiardtable	pool table	Fail	Synonym	Accept	Accepted as equivalent.
9	bird	bird	Pass	–	–	–
10	boat	boat	Pass	–	–	–
11	bowl	bowl	Pass	–	–	–
12	butterfly	butterfly	Pass	–	–	–
13	camel	camel	Pass	–	–	–
14	car	car	Pass	–	–	–
15	cat	cat	Pass	–	–	–
16	chook	rooster	Fail	Synonym	Accept	Chook is Australian English for chicken.
17	coffeecup	cup	Pass	–	–	–
18	cow	cow	Pass	–	–	–
19	deer	antelope	Fail	Wrong†	Error†	Originally marked as related but imprecise; rejected in final review.
20	dog	dog	Pass	–	–	–
21	frenchhorn	french horn	Fail	Variant	Accept	Equivalent after word separation.

Continued on the next page.

Table 10 (continued): CoSal2015

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
22	frog	frog	Pass	–	–	–
23	goldenfish	fish	Pass	–	–	–
24	guitar	guitar	Pass	–	–	–
25	hammer	hammer	Pass	–	–	–
26	helmet	helmet	Pass	–	–	–
27	horse	horse	Pass	–	–	–
28	ladybird	ladybug	Fail	Synonym	Accept	British/US English equivalents.
29	lemon	lemon	Pass	–	–	–
30	lizard	gecko	Fail	Synonym	Accept	A gecko is a type of lizard.
31	mobulidae	ray	Fail	Synonym	Accept	Mobulidae denotes the manta/devil-ray family.
32	monkey	monkey	Pass	–	–	–
33	motorbike	motorcycle	Fail	Synonym	Accept	Accepted as equivalent.
34	mouse	mouse	Pass	–	–	–
35	mushroom	mushroom	Pass	–	–	–
36	penguin	penguin	Pass	–	–	–
37	pepper	pepper	Pass	–	–	–
38	piano	piano	Pass	–	–	–
39	pig	pig	Pass	–	–	–
40	pineapple	pineapple	Pass	–	–	–
41	rabbit	hare	Fail	Synonym	Accept	Accepted as a related name.
42	sealion	seal	Pass	–	–	–
43	snail	snail	Pass	–	–	–
44	snake	snake	Pass	–	–	–
45	sofa	couch	Fail	Synonym	Accept	Accepted as equivalent.
46	starfish	starfish	Pass	–	–	–
47	tiger	tiger	Pass	–	–	–
48	train	train	Pass	–	–	–
49	turtle	turtle	Pass	–	–	–
50	viola	cello	Fail	Wrong	Error	Misidentifies the instrument.

Table 10: Complete class-wise consensus predictions and review annotations on CoSal2015 (50 classes). Type and Review follow the definitions in Section A.3. The † marks the final adjudication of *deer→antelope*, counted as an error.

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
1	airplane	airplane	Pass	–	–	–
2	ant	ant	Pass	–	–	–
3	antelope	antelope	Pass	–	–	–
4	apple	apple	Pass	–	–	–
5	armadillo	armadillo	Pass	–	–	–
6	artichoke	artichoke	Pass	–	–	–
7	axe	axe	Pass	–	–	–
8	baby_bed	bed	Pass	–	–	–
9	bagel	bagel	Pass	–	–	–
10	banana	banana	Pass	–	–	–
11	band_aid	bandage	Fail	Synonym	Accept	Band-aid is accepted as bandage.
12	banjo	banjo	Pass	–	–	–
13	baseball	baseball	Pass	–	–	–
14	basketball	basketball	Pass	–	–	–
15	beaker	egg	Fail	Wrong	Error	Misidentifies the object.

Continued on the next page.

Table 11 (continued): CoSOD3k

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
16	bear	bear	Pass	–	–	–
17	bee	bee	Pass	–	–	–
18	bell_pepper	pepper	Pass	–	–	–
19	bench	bench	Pass	–	–	–
20	bicycle	bicycle	Pass	–	–	–
21	binder	notebook	Fail	Near-syn.	Partial	Related, but imprecise.
22	bird	bird	Pass	–	–	–
23	bookshelf	book	Pass	–	–	–
24	bow	archery bow	Pass	–	–	–
25	bow_tie	bow tie	Fail	Format	Accept	Underscore in GT; space in prediction.
26	bowl	bowl	Pass	–	–	–
27	brassiere	bra	Pass	–	–	–
28	bus	bus	Pass	–	–	–
29	butterfly	butterfly	Pass	–	–	–
30	camel	camel	Pass	–	–	–
31	can_opener	can opener	Fail	Format	Accept	Accepted as equivalent.
32	car	car	Pass	–	–	–
33	cart	cart	Pass	–	–	–
34	cattle	cow	Fail	Synonym	Accept	Accepted as equivalent.
35	cello	cello	Pass	–	–	–
36	centipede	centipede	Pass	–	–	–
37	chain_saw	chainsaw	Fail	Variant	Accept	Accepted as equivalent.
38	chime	bell	Fail	Near-syn.	Partial	Related, but imprecise.
39	cocktail_shaker	shaker	Pass	–	–	–
40	computer_keyboard	keyboard	Pass	–	–	–
41	computer_mouse	mouse	Pass	–	–	–
42	cream	bottle	Fail	Wrong	Error	Names the container instead of its contents.
43	croquet_ball	ball	Pass	–	–	–
44	crutch	crutches	Pass	–	–	–
45	cucumber	cucumber	Pass	–	–	–
46	cup_or_mug	cup	Pass	–	–	–
47	digital_clock	clock	Pass	–	–	–
48	dog	dog	Pass	–	–	–
49	domestic_cat	cat	Pass	–	–	–
50	dragonfly	dragonfly	Pass	–	–	–
51	drum	drum	Pass	–	–	–
52	dumbbell	dumbbell	Pass	–	–	–
53	electric_fan	fan	Pass	–	–	–
54	elephant	elephant	Pass	–	–	–
55	face_powder	cosmetics	Fail	Hypernym	Partial	Too broad.
56	fig	fig	Pass	–	–	–
57	flower_pot	plant	Fail	Hypernym	Partial	Too broad.
58	flute	flute	Pass	–	–	–
59	fox	fox	Pass	–	–	–
60	french_horn	french horn	Fail	Format	Accept	Accepted as equivalent.
61	frog	frog	Pass	–	–	–
62	frying_pan	pan	Pass	–	–	–
63	giant_panda	panda	Pass	–	–	–
64	goldfish	fish	Pass	–	–	–
65	golf_ball	golf ball	Fail	Format	Accept	Accepted as equivalent.

Continued on the next page.

Table 11 (continued): CoSOD3k

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
66	golfcart	golf cart	Fail	Variant	Accept	Accepted as equivalent.
67	guitar	guitar	Pass	–	–	–
68	hair_dryer	hair dryer	Fail	Format	Accept	Accepted as equivalent.
69	hair_spray	hair spray	Fail	Format	Accept	Accepted as equivalent.
70	hamburger	hamburger	Pass	–	–	–
71	hammer	hammer	Pass	–	–	–
72	hamster	hamster	Pass	–	–	–
73	harmonica	harmonica	Pass	–	–	–
74	harp	harp	Pass	–	–	–
75	hat_with_a_wide_brim	hat	Pass	–	–	–
76	head_cabbage	cabbage	Pass	–	–	–
77	helmet	helmet	Pass	–	–	–
78	hippopotamus	hippo	Pass	–	–	–
79	horse	horse	Pass	–	–	–
80	iPod	ipod	Pass	–	–	–
81	isopod	isopod	Pass	–	–	–
82	jellyfish	jellyfish	Pass	–	–	–
83	koala_bear	koala	Pass	–	–	–
84	ladybug	ladybug	Pass	–	–	–
85	lamp	lamp	Pass	–	–	–
86	laptop	laptop	Pass	–	–	–
87	lemon	lemon	Pass	–	–	–
88	lion	lion	Pass	–	–	–
89	lipstick	lipstick	Pass	–	–	–
90	lizard	lizard	Pass	–	–	–
91	lobster	lobster	Pass	–	–	–
92	maraca	maracas	Pass	–	–	–
93	microwave	microwave	Pass	–	–	–
94	milk_can	milk can	Fail	Format	Accept	Accepted as equivalent.
95	monkey	monkey	Pass	–	–	–
96	motorcycle	motorcycle	Pass	–	–	–
97	mushroom	mushroom	Pass	–	–	–
98	nail	nail	Pass	–	–	–
99	oboe	flute	Fail	Wrong	Error	Misidentifies the instrument.
100	orange	orange	Pass	–	–	–
101	otter	otter	Pass	–	–	–
102	pencil_box	pencil case	Fail	Synonym	Accept	Accepted as equivalent.
103	perfume	perfume	Pass	–	–	–
104	person	guitar	Fail	Wrong	Error	Incorrect object identity.
105	piano	piano	Pass	–	–	–
106	pineapple	pineapple	Pass	–	–	–
107	pitcher	jar	Fail	Near-syn.	Partial	Related, but imprecise.
108	pizza	pizza	Pass	–	–	–
109	pomegranate	pomegranate	Pass	–	–	–
110	popsicle	popsicle	Pass	–	–	–
111	porcupine	porcupine	Pass	–	–	–
112	power_drill	drill	Pass	–	–	–
113	pretzel	pretzel	Pass	–	–	–
114	punching_bag	sandbag	Fail	Near-syn.	Partial	Related, but imprecise.
115	purse	bag	Fail	Hypernym	Partial	Too broad.
116	rabbit	rabbit	Pass	–	–	–
117	ray	rayfish	Pass	–	–	–
118	red_panda	red panda	Fail	Format	Accept	Accepted as equivalent.

Continued on the next page.

Table 11 (continued): CoSOD3k

No.	GT class	Predicted c^*	Strict	Type	Review	Explanation
119	refrigerator	refrigerator	Pass	–	–	–
120	remote_control	remote	Pass	–	–	–
121	rubber_eraser	pencil	Fail	Wrong	Error	Misidentifies the object.
122	rugby_ball	ball	Pass	–	–	–
123	salt_or_pepper_shaker	shaker	Pass	–	–	–
124	scorpion	scorpion	Pass	–	–	–
125	screwdriver	screwdriver	Pass	–	–	–
126	seal	seal	Pass	–	–	–
127	sheep	ram	Fail	Synonym	Accept	A ram is a male sheep.
128	skunk	skunk	Pass	–	–	–
129	snail	snail	Pass	–	–	–
130	snake	snake	Pass	–	–	–
131	snowmobile	snowmobile	Pass	–	–	–
132	snowplow	snowplow	Pass	–	–	–
133	soap_dispenser	soap	Pass	–	–	–
134	soccer_ball	ball	Pass	–	–	–
135	sofa	couch	Fail	Synonym	Accept	Accepted as equivalent.
136	spatula	spatula	Pass	–	–	–
137	squirrel	squirrel	Pass	–	–	–
138	starfish	starfish	Pass	–	–	–
139	stethoscope	stethoscope	Pass	–	–	–
140	strainer	strainer	Pass	–	–	–
141	strawberry	strawberry	Pass	–	–	–
142	swine	pig	Fail	Synonym	Accept	Accepted as equivalent.
143	syringe	syringe	Pass	–	–	–
144	tape_player	speaker	Fail	Wrong	Error	Misidentifies the object.
145	tennis_ball	tennis ball	Fail	Format	Accept	Accepted as equivalent.
146	tick	tick	Pass	–	–	–
147	tie	tie	Pass	–	–	–
148	tiger	tiger	Pass	–	–	–
149	toaster	toaster	Pass	–	–	–
150	traffic_light	traffic light	Fail	Format	Accept	Accepted as equivalent.
151	train	train	Pass	–	–	–
152	trumpet	trumpet	Pass	–	–	–
153	turtle	turtle	Pass	–	–	–
154	violin	violin	Pass	–	–	–
155	washer	washer	Pass	–	–	–
156	water_bottle	bottle	Pass	–	–	–
157	watercraft	boat	Fail	Hypernym	Partial	Too broad.
158	whale	whale	Pass	–	–	–
159	wine_bottle	wine	Pass	–	–	–
160	zebra	zebra	Pass	–	–	–

Table 11: Complete class-wise consensus predictions and review annotations on CoSOD3k (160 classes). Type and Review follow the definitions in Section A.3.

B Additional Qualitative Results

Figure 6 presents additional CoSEC predictions on twelve CoCA categories that do not appear in Figure 4. The three columns examine clutter and distractors, appearance variation, and thin or small targets. Each column stacks four distinct categories with four images per category. For each category, the Input, ground-truth (GT), and CoSEC rows use the same image order; white pixels denote foreground in the two mask rows.

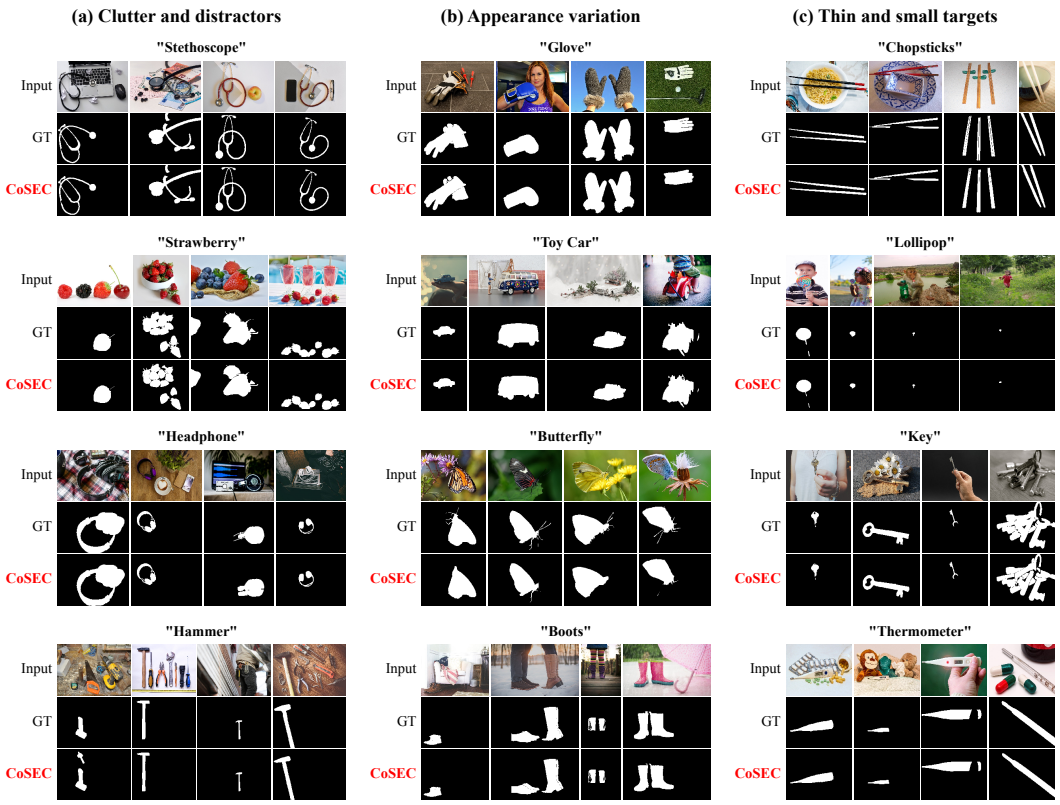


Figure 6: Additional CoSEC visualizations on CoCA. Each of the twelve categories contains four examples, shown as aligned Input, ground-truth (GT), and CoSEC rows (white denotes foreground). The columns show (a) clutter and distractors: stethoscopes, strawberries, headphones, and hammers; (b) appearance variation: gloves, toy cars, butterflies, and boots; and (c) thin and small targets: chopsticks, lollipops, keys, and thermometers.

Clutter and distractors. In panel (a), stethoscopes appear alongside other medical equipment, strawberries among other fruit and food, and headphones and hammers near related accessories or tools. These scenes contain several plausible foreground objects. CoSEC generally assigns its mask to the category shared by the image group while leaving the surrounding objects in the background.

Appearance variation. Panel (b) spans gloves, toy cars, butterflies, and boots with substantial changes in color, shape, scale, and viewpoint. Despite these differences between images of the same category, the CoSEC masks remain focused on the shared object and broadly follow the GT regions. This illustrates how an explicit category-level consensus can guide localization across varied appearances.

Thin and small targets. In panel (c), chopsticks and thermometers have narrow structures, whereas lollipops and keys can occupy only a small fraction of the image. CoSEC recovers many of these slender or small regions even when larger objects dominate the scene. Differences from GT are still visible at some fine boundaries and tiny details, which remain difficult to segment precisely.

C Single-Image Consensus Estimation

To further evaluate the effectiveness of the explicit semantic consensus framework across different consensus estimation procedures, we consider generating c^* from independent single-image VLM queries. Its principal difference from the group-input method in Section 3.2 lies in the generation of c^*; the subsequent CGS follows exactly the same procedure as in Section 3.3. We therefore focus in this appendix on generating c^* from single-image inputs and evaluating the number of candidate prompts used in this process.

C.1 METHOD

The key challenge in generating c^* through independent single-image queries, compared with joint group-level reasoning, is the absence of cross-image context within each VLM query. From one image alone, the VLM cannot directly determine which salient object also appears in the other group images. A locally prominent distractor may therefore be selected even when the shared target is salient. Moreover, restricting the VLM to a single object name may prematurely exclude the true shared target before cross-image evidence can be considered. To preserve alternatives, we first generate a short caption describing the main objects, then retain multiple candidate names for group-level voting.

Figure 7 illustrates the generation of explicit semantic consensus c^* from independent single-image inputs. Each image is captioned separately, and up to k candidate object names are retained from its caption. Cross-image voting then selects the object name receiving the most support across individual images. This allows a shared target to contribute to the consensus even when it is not the highest-ranked candidate in every image.

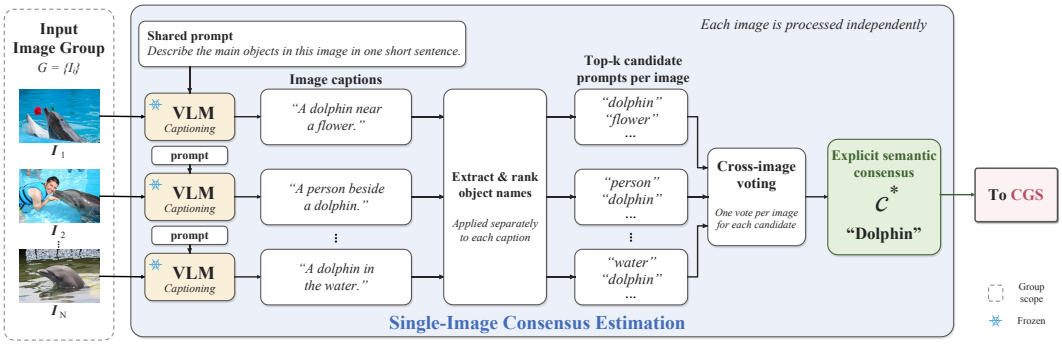


Figure 7: Single-image consensus estimation. A frozen VLM independently captions each image using a shared prompt. Object-name extraction and ranking operate separately on each caption, retaining at most k candidate prompts per image. Each image contributes at most one vote to each retained candidate, and the candidate supported by the most images becomes the explicit semantic consensus c^*. CGS then uses this consensus as in the main paper. Captions and candidate names are illustrative.

Independent image captioning. Following the problem description in the main paper, we consider an image group $\mathcal{G} = \{I_i\}_{i=1}^N$ and use a frozen VLM $\mathcal{V}$. For each image I_i, we generate a caption using the same fixed instruction p_{single}:

$$t_i = \mathcal{V}(I_i; p_{\text{single}}), \tag{6}$$

where t_i is a short caption describing the objects in I_i. The same instruction is shared for all the images:

$$p_{\text{single}} : \textit{Describe the main objects in this image in one short sentence.}$$

Each query contains only one image and provides no predefined category names or ground-truth annotations. The instruction p_{single} guides the VLM to produce the descriptive text t_i, which is then used to obtain candidate object names.

Per-image candidate object extraction. The caption t_i may describe multiple objects and their surroundings, leaving the target identity ambiguous for the subsequent CGS, which requires a concise object name as guidance. We therefore convert each caption into candidate object names for subsequent cross-image voting. Specifically, we tokenize t_i and feed the token sequence into a frozen BERT Devlin et al. (2019) model to obtain contextualized token representations. We then reconstruct words by merging their constituent subwords, filter out stopwords and words ending in "ing" or "ed". After lowercasing, trimming surrounding whitespace, and deduplication, the remaining words form the candidate names $\{u_{i,j}\}_{j=1}^{m_i}$, where m_i is the number of distinct retained words.

We rank these candidates by their semantic relevance to the caption. Let $\mathcal{B}(u_{i,j}; t_i)$ denote the average final-layer representations of the subword tokens corresponding to candidate $u_{i,j}$ within the

encoded caption t_i, and let $\mathcal{B}(t_i)$ denote the average over all valid caption tokens, excluding special and padding tokens. Both representations are obtained from the same BERT encoding of t_i. The relevance score is:

$$s_{i,j} = \frac{\mathcal{B}(u_{i,j}; t_i)^\top \mathcal{B}(t_i)}{\|\mathcal{B}(u_{i,j}; t_i)\|_2 \|\mathcal{B}(t_i)\|_2}, \tag{7}$$

We sort and reindex the candidate names by decreasing $s_{i,j}$, yielding the ordered list $\mathcal{R}(t_i) = (u_{i,1}, \ldots, u_{i,m_i})$. To preserve alternative target identities, we retain at most k candidates per image:

$$\mathcal{C}_i^{(k)} = \{u_{i,j}\}_{j=1}^{\min(k,m_i)}. \tag{8}$$

Here, k controls the maximum number of candidate prompts contributed by each image; all candidates are retained when $m_i < k$. Extraction and ranking are performed independently for each caption, and evidence from different images is combined only in the subsequent voting step.

Cross-image voting. To identify the shared object among the candidates retained from individual images, we aggregate their support through cross-image voting. Let $\mathcal{C} = \bigcup_{i=1}^{N} \mathcal{C}_i^{(k)}$ denote the union of the per-image candidate sets. We count the frequency of each candidate name exists in $\mathcal{C}$ and select the most frequent one as the explicit semantic consensus, formulated as:

$$c^* = \arg\max_{c \in \mathcal{C}} \sum_{i=1}^{N} \mathbb{1}\left[c \in \mathcal{C}_i^{(k)}\right], \tag{9}$$

where $\mathbb{1}[\cdot]$ is the indicator function. Each image contributes at most one vote to a given candidate, regardless of how often the name occurs in its caption. Increasing k allows additional object names from each image to participate in the vote, potentially recovering a shared target that ranks below an image-specific distractor. The resulting c^* remains a compact textual concept. Its selection uses caption-derived candidates alone, without SAM 3 masks or confidence scores.

Consensus-guided segmentation. Once c^* is selected, subsequent per-image segmentation follows the same CGS procedure described in Section 3.3.

C.2 Ablation on the Number of Candidate Prompts

	CoCA				CoSal2015				CoSOD3k			
Setting	MAE↓	$F_\beta^{\max}$ ↑	$E^{\max}$ ↑	S_α ↑	MAE↓	$F_\beta^{\max}$ ↑	$E^{\max}$ ↑	S_α ↑	MAE↓	$F_\beta^{\max}$ ↑	$E^{\max}$ ↑	S_α ↑
TF-SSD	0.077	0.686	0.815	0.763	0.089	0.899	0.926	0.890	0.090	0.860	0.908	0.861
CoSEC (default)	**0.023**	0.885	**0.940**	**0.909**	**0.020**	**0.953**	**0.963**	**0.933**	**0.040**	0.874	0.912	**0.885**
k=1	0.0417	0.7942	0.8776	0.8518	0.0693	0.8586	0.8862	0.8617	0.1356	0.6413	0.7441	0.7228
k=2	0.0386	0.8405	0.9091	0.8748	0.0332	0.9323	0.9435	0.9156	0.0761	0.8162	0.8697	0.8367
k=3	0.0342	0.8601	0.9204	0.8854	0.0207	0.9510	0.9594	0.9313	0.0639	0.8408	0.8898	0.8555
k=4	0.0315	0.8655	0.9241	0.8896	0.0206	0.9515	0.9598	0.9316	0.0622	0.8496	0.8945	0.8605
k=5	0.0305	0.8843	0.9357	0.8993	0.0206	0.9515	0.9598	0.9316	0.0509	0.8710	0.9113	0.8768
k=6	0.0266	**0.8923**	0.9379	0.9056	0.0212	0.9485	0.9577	0.9296	0.0493	**0.8773**	**0.9162**	0.8806
k=7	0.0301	0.8818	0.9328	0.8993	0.0212	0.9485	0.9577	0.9296	0.0537	0.8752	0.9134	0.8773
k=8	0.0301	0.8818	0.9328	0.8993	0.0260	0.9399	0.9481	0.9212	0.0535	0.8765	0.9145	0.8781

Table 12: Ablation on the number of candidate prompts per image for BLIP-2. TF-SSD and default CoSEC (joint Qwen group input) are reference rows reproduced from Table 1 at their reported precision. Bold denotes the best value across all rows for each dataset and metric, including both reference methods; ties are all highlighted.

We use BLIP-2 Li et al. (2023) as the frozen captioning VLM and vary k from 1 to 8 to examine the effect of the number of candidate prompts retained per image. The captioning and candidate-ranking procedures remain fixed, while k determines how many candidates from each image participate in cross-image voting. The resulting consensus c^* guides the same CGS for all settings. Table 12 reports the four evaluation metrics on all three benchmarks, together with TF-SSD and our default CoSEC with joint Qwen group input for comparison.

Effect of the candidate count. Retaining multiple candidate names improves consensus-guided segmentation over k=1. At k=6, $F_\beta^{\max}$ increases from 0.7942 to 0.8923 on CoCA, from 0.8586 to

0.9485 on CoSal2015, and from 0.6413 to 0.8773 on CoSOD3k. Performance is relatively stable within k=5→8: the ranges of $F_\beta^{\max}$ are $0.8818 \rightarrow 0.8923$, $0.9399 \rightarrow 0.9515$, and $0.8710 \rightarrow 0.8773$, respectively. The variation is therefore at most 0.0116 on any benchmark, showing that strong performance extends across several candidate counts.

Comparison with existing methods. For every $k \in \{5, 6, 7, 8\}$, the BLIP-2 single-image variant outperforms TF-SSD He et al. (2026) on all four metrics across all three benchmarks. Its $F_\beta^{\max}$ also exceeds the strongest prior scores reported in Table 1 (0.686, 0.920, and 0.868, respectively) throughout this range. Thus, the advantage holds across a shared range of candidate counts without choosing a separate optimal k for each benchmark.

Comparison with default CoSEC. At a common k=6, the BLIP-2 single-image variant achieves $F_\beta^{\max}$ scores of 0.8923, 0.9485, and 0.8773, close to default CoSEC's 0.885, 0.953, and 0.874 with joint Qwen2.5-VL Bai et al. (2025) group input. The absolute differences are no larger than 0.0073, with the single-image variant scoring higher on CoCA and CoSOD3k. Default CoSEC retains stronger MAE and S_α across the three benchmarks. Overall, these results show that independent captioning and textual voting can retain competitive segmentation performance through the same explicit-consensus interface, even with a different VLM and input format.