V^2-SAM: Marrying SAM2 with Multi-Prompt Experts for Cross-View Object Correspondence

Supplementary Material

Contents

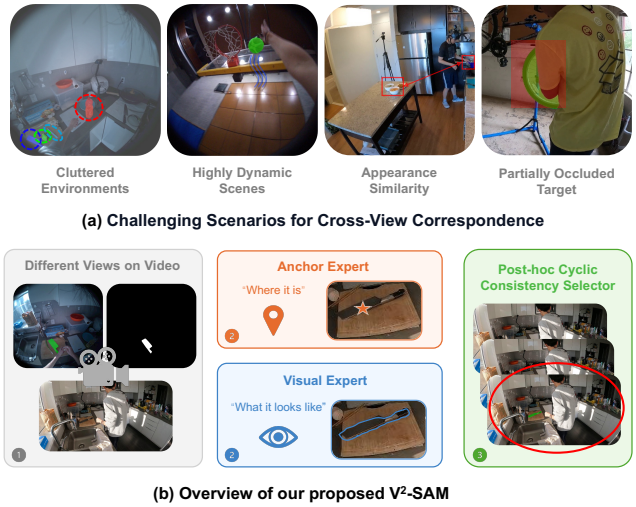


Figure 7. Cross-view correspondence challenges and our method.

1. More Related Work

1.1. Segment Anything Model

The Segment Anything Model (SAM) is a prompt-driven foundation model for universal image localization [26, 27, 40] and segmentation, capable of producing high-quality masks from simple inputs like points or bounding boxes. It has inspired domain-specific extensions such as MedSAM [66] for medical imaging, InstructSAM [65] for remote sensing, and video or language-aware variants that enhance temporal and semantic reasoning across diverse tasks.

SAM2 [44] is one of the most powerful segmentation models to date, demonstrating outstanding generalization ability in images and videos. It consists of a *Image Encoder* for visual feature extraction, a *Prompt Encoder* for embedding points, boxes, or masks, and a *Mask Decoder* that predicts object masks using features from both encoders. A memory mechanism further enables temporal mask propagation across video frames. However, since all prompts are defined within the coordinate system of the target image, SAM2 is inherently non-trivial to adapt to cross-view scenarios, where object position, scale, shape, and appearance often vary drastically across views.

Our framework is built upon SAM2, retaining its core components. Specifically, we keep the *SAM2 Encoder* $\phi(\cdot)$, *Prompt Encoder*, and *Mask Decoder*, while discarding memory-related modules to focus on frame-level correspondence. This design allows the framework to generalize seamlessly across both cross-view image and video tasks.

To unlock SAM2's potential for cross-view object correspondence, we introduce four novel modules: 1) a *Cross-View Anchor Prompt Generator (V^2-Anchor)* that transfers the query mask's spatial information to the target view using DINOv3 $\varphi(\cdot)$'s geometry-aware feature space, for the first time enabling coordinate-based prompting across views; 2) a *Cross-View Visual Prompt Generator (V^2-Visual)* that leverages object appearance cues and refines them through a learnable mapping between views; 3) a *Multi-Expert Training* mechanism that jointly learns spatial, visual, and fused experts for complementary reasoning; and 4) a *Post-hoc Cyclic Consistency Selector (PCCS)* that adaptively selects the most reliable expert at inference based on cross-view mask consistency. Together, these components form our V^2-SAM, a unified segmentation framework that bridges spatial alignment and semantic association across drastically different viewpoints.

1.2. Mixture-of-Experts in Vision

Recent advances in the Mixture-of-Experts [15] (MoE) paradigm have demonstrated strong potential for scalable and adaptive visual modeling via input-dependent expert routing in computer vision [40, 68]. Building on this idea, TimeExpert [56] extends dynamic routing to spatiotemporal modeling, FlexTrack [49] introduces heterogeneous experts for adaptive computation, and SM3Det [22] employs sparse grid-level experts for multi-modal detection for remote sensing [28, 37, 39]; meanwhile, XTrack [50] and Pro-

Dataset	Subset	Split	Pairs	Masks	# Classes
Ego-Exo4D	Ego2Exo	Train	110K	523K	∼28
	Ego2Exo	Test	41K	200K	∼35
	Exo2Ego	Train	123K	567K	∼29
	Exo2Ego	Test	47K	219K	∼35
HANDAL-X	-	Train	39K	78K	17
	-	Test	13K	26K	17
DAVIS-17	-	Train	2.8K	12.7K	8
	-	Test	1.4K	5.3K	5

Table 6. Statistics of datasets used in our experiments. Ego-Exo4D is divided into directional subsets Ego2Exo and Exo2Ego.

Hyperparameters	Value
Batch size (per device)	16
Gradient accumulation steps	4
Effective batch size	64
Training epochs	24
Validation frequency	every 2 epochs
Optimizer	AdamW
Learning rate	4×10^{-5}
Adam betas	$(0.9, 0.999)$
Weight decay	0.05
Gradient clipping norm	1.0
Precision	bfloat16 (mixed precision)
Warm-up strategy	linear warm-up
Warm-up ratio	0.05 of total epochs
Learning rate schedule	cosine annealing

Table 7. Training hyperparameters used in our experiments.

MoE [53] further improve modality fusion and routing precision. Despite these advances, existing MoE designs primarily focus on single-view or modality-level adaptation. In contrast, we propose a multi-prompt expert framework that adaptively selects spatial and visual experts to handle cross-view correspondence under drastic viewpoint and appearance changes.

2. Challenges in Cross-View Object Correspondence

Cross-view object correspondence in real-world environments remains highly challenging due to substantial intra-scene variations and visual ambiguity across viewpoints, as shown in Fig. 7. First, *cluttered scenes* with numerous overlapping objects introduce significant distractors, making it difficult to reliably localize the same instance across views. Second, *highly dynamic settings*, where either the camera or objects exhibit rapid motion, lead to drastic changes in appearance, lighting, and geometry. Third, the presence of *appearance-similar objects* (e.g., tools with comparable shapes and textures) often results in ambiguous matching. Finally, *partial occlusions* may cause the target to be only partially visible or temporarily absent in certain views, further increasing correspondence difficulty.

To address these challenges, we propose V^2-SAM, a unified framework that jointly leverages (i) a *Anchor Expert* to reason about object location ("where it is"), (ii) a *Visual Expert* to capture fine-grained appearance cues ("what it looks like"), and (iii) a *Post-hoc Cyclic Consistency Selector* to enforce cross-view agreement. Collectively, these components enable more robust and accurate cross-view correspondence in complex, real-world settings.

3. More Implementation Details

3.1. Dataset Settings

Tab. 6 provides a quantitative overview of the datasets used in our experiments. Our primary supervision comes from Ego-Exo4D, where we leverage two directional splits: *Ego2Exo* and *Exo2Ego*. Each direction includes both training and testing sets, totaling over 320K pairs and 1.5M masks across roughly 30 semantic categories. This large-scale paired data enables cross-view correspondence learning between egocentric and exocentric perspectives. HANDAL-X contributes an additional 52K pairs and 104K masks covering 17 categories of manipulable objects, providing object-level diversity while still maintaining a focus on real-world, robotics-relevant items. Finally, DAVIS-17 provides 4.2K pairs and 18K high-quality masks for multi-object video segmentation, with 8 training classes and 5 testing classes, serving as a benchmark for generalization in dense video segmentation.

Ego-Exo4D. Ego-Exo4D contains synchronized egocentric and exocentric videos of human activities. We use two directional subsets: Ego2Exo (110K/41K train/test pairs, 523K/200K masks) and Exo2Ego (123K/47K train/test pairs, 567K/219K masks), spanning roughly 28–35 classes. These splits provide large-scale supervision for cross-view mapping.

HANDAL-X. HANDAL-X contains real-world manipulable objects with 6-DoF pose labels. We use both the train (39K pairs, 78K masks) and test (13K pairs, 26K masks) splits, spanning 17 object categories. Its focus on object instances rather than semantic classes makes it well suited for object-centric manipulation tasks.

DAVIS-2017. DAVIS-2017 serves as a high-quality benchmark for multi-object video segmentation. We use 2.8K training pairs (12.7K masks, 8 classes) and 1.4K testing pairs (5.3K masks, 5 classes), providing dense pixel-level masks for evaluating generalization.

3.2. Training Hyperparameters

A summary of our training hyperparameters is provided in Tab. 7. We train our model for 24 epochs with an effective batch size of 64, obtained by using a per-device batch size of 16 and four gradient accumulation steps. The optimization is performed using the AdamW optimizer with a learning rate of 4×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.05. We employ mixed-precision training with `bfloat16` and dynamic loss scaling, and apply gradient clipping with a maximum norm of 1.0 to stabilize training. The learning rate follows a linear warm-up schedule for the first 5% of epochs, followed by a cosine annealing policy with a minimum learning rate of zero. Validation is conducted every two epochs.

3.3. Model settings

The model configuration is summarized in Tab. 8. Our model is built upon the V^{2}SAM framework with a fully trainable SAM2 decoder and a grounding encoder. We initialize the model with pretrained weights and resize all input images to 1024 pixels using a direct resizing operator. The training objective combines a sigmoid-activated binary cross-entropy loss and a naive Dice loss, weighted by 2.0 and 0.5 respectively. We additionally employ point-sampled supervision to enhance mask quality. The entire system is trained using a length-grouped sampler and a video-aware collation strategy to accommodate variable-length multimodal data.

Model Settings	Value
Backbone model	SAM2
Decoder	SAM2 decoder (trainable)
Encoder 1	SAM2 encoder
Encoder 2	DINOv3 encoder
Pretrained checkpoint	pretrained SAM2
Image pre-processing	direct resizing to 1024 px
Use point-sampled supervision	Yes
Binary Cross-Entropy loss weight	2.0
Dice loss weight	0.5
Dice activation	sigmoid + activation
Dice formulation	naive Dice variant

Table 8. Model configuration and loss functions.

4. More Experiments

4.1. Ablation on Submodule

Tab. 11 presents the ablation results of the proposed components, including the two submodules of V^2-Visual (Semantic Mapping and Spatial Mapping), the associated losses $\mathcal{L}_v$ and $\mathcal{L}_s$, and the V^2-Anchor. Each component contributes positively to overall performance, while V^2-Anchor yields the greatest improvement.

Comparing the first two rows, the configuration with Semantic Mapping and $\mathcal{L}_v$ slightly outperforms the one using Spatial Mapping and $\mathcal{L}_s$, suggesting that enforcing semantic consistency plays a more crucial role than spatial alignment when transferring visual cues across views. When both mapping modules are jointly enabled with their respective losses (third row), we observe further gains in overall IoU, demonstrating that semantic and spatial constraints are complementary and collaboratively enhance cross-view coherence. The most substantial improvement is observed after introducing V^2-Anchor (fourth row). Acting as a stable cross-view reference, V^2-Anchor significantly strengthens ego–exo alignment, leading to relative improvements of 30.2% (Ego2Exo), 4.7% (Exo2Ego), and 15.7% (Total). These results indicate that anchoring the visual prompts effectively reduces cross-view ambiguity and stabilizes the mapping process.

Overall, the ablation study highlights the synergy among the proposed components: the V^2-Visual module establishes the foundation for consistent cross-view representation, while V^2-Anchor further amplifies this effect, enabling the full framework to achieve the best performance.

4.2. Ablation on V^2-Anchor

Tab. 9 reports the impact of varying the number of sparse anchor points in the V^2-Anchor Expert. In the Ego→Exo setting, performance monotonically degrades as anchors become denser, dropping from 38.7 IoU with a single anchor to 32.2, 28.7, and 21.8 with 5, 10, and 30 anchors. A similar trend is observed in the Exo→Ego direction, where the 1-anchor configuration achieves the best result (41.6 IoU), while denser anchors significantly reduce accuracy (e.g., 18.3 and 18.5 IoU with 10 and 30 anchors).

Setting	1 pt	5 pts	10 pts	30 pts
Ego→Exo	38.7	32.2	28.7	21.8
Exo→Ego	41.6	26.1	18.3	18.5

Table 9. Ablation of sparse anchor point count in the V^2-Anchor Expert. We report IoU under Ego→Exo and Exo→Ego settings using 1, 5, 10, and 30 correspondence points.

These results suggest that a minimal set of high-confidence cross-view correspondences provides stronger supervisory constraints, whereas denser anchor distributions may introduce view-specific noise or limit generalization. Overall, sparse anchoring proves more effective for cross-view alignment across both transfer directions.

4.3. Ablation on the PCCS

Tab. 10 compares our PCCS with the prior Cycle-Mask Selector, which reconstructs query-view masks for cyclic

Decoders	Selector	Ego2Exo IoU↑	Exo2Ego IoU↑	Runtime (ms/sample)↓	FLOPs (G/sample)↓
A+B	Cycle-Mask (Prior)	42.60	46.73	620	2188.58 GFLOPs
	Cycle-Points (Ours)	**42.71**	**48.17**	**510**	**2173.58 GFLOPs**
A+B+C	Cycle-Mask (Prior)	46.27	49.43	820	2207.13 GFLOPs
	Cycle-Points (Ours)	**46.31**	**49.61**	**760**	**2184.63 GFLOPs**

Table 10. Ablation on the Post-hoc Cyclic Consistency Selector in **Ego↔Exo** correspondence. We compare **Cycle-Points (Ours)** with the mask-based **Cycle-Mask (Prior)** on two decoder combinations: A+B (Anchor+Visual) and A+B+C (Anchor+Visual+Fusion).

V^2-Visual Semantic Mapping	V^2-Visual Spatial Mapping	$\mathcal{L}_v$	$\mathcal{L}_s$	V^2-Anchor	Ego2Exo IoU↑	Exo2Ego IoU↑	Total IoU↑
✓	✓	-	✓	-	34.19	45.16	39.68
✓	-	✓	-	-	34.98	46.46	40.72
✓	✓	✓	✓	-	36.17	46.63	41.40
✓	✓	✓	✓	✓	44.51	47.29	45.90
Relative Gain % of x with respect to y $\frac{(x-y)}{y}$					+30.2%	+4.7%	+15.7%

Table 11. Ablation study of the proposed modules V^2-Visual, V^2-Anchor, $\mathcal{L}_v$ and $\mathcal{L}_s$ on the test set. The Cross-View Visual Prompt Generator (V^2-Visual) consists of two submodules, Semantic Mapping and Spatial Mapping, which correspond to Semantic Constraint and Spatial Constraint, respectively.

validation. Instead of predicting masks, our method directly measures geometric consistency between predicted query points and sampled reference points from the raw prompt mask, selecting the expert with the smallest cyclic distance. Despite its simpler design, Cycle-Points achieves comparable or even higher accuracy while notably reducing computation. For the two-expert setup (A+B), it improves Exo→Ego IoU by +1.4 and shortens runtime by 110 ms per sample. With three experts (A+B+C), it maintains accuracy while reducing both latency and FLOPs.

These results demonstrate that expert selection can be effectively driven by point-level cyclic consistency alone, providing a lightweight yet equally reliable alternative to mask-based cyclic reasoning.

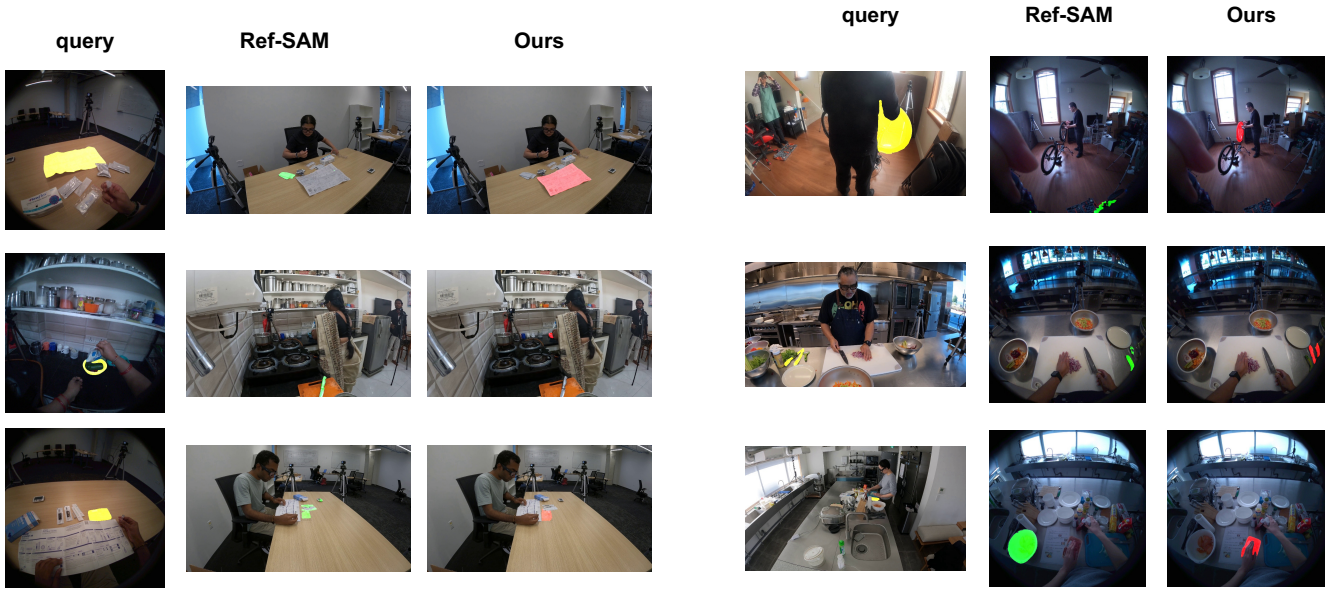
4.4. More Visual Analytics.

Ref-SAM VS. V^2-SAM. Fig. 8a and Fig. 8b show the visual quality comparison of Ref-SAM and our proposed V^2-SAM. Ref-SAM often fails to accurately localize target objects in cluttered scenes or when multiple similar objects are present. In contrast, our method generates precise and consistent predictions across both directions, illustrating superior robustness and cross-view generalization. These results further confirm that our method effectively bridges the viewpoint gap between egocentric and exocentric observations.

Different Experts. Fig. 9 and Fig. 10 present the comparative results obtained from multiple experts and the PCCS for the Ego2Exo and Exo2Ego tasks, respectively. The results demonstrate that combining expert assessments with the PCCS has a substantial influence on the final selection process. Moreover, different experts exhibit complementary strengths in understanding distinct scenario types, underscoring the benefit of aggregating diverse human expertise with system-level decision-making.

PCCS Consensus Analysis. Fig. 11 and Fig. 12 analyze the mechanism of PCCS under the Ego2Exo and Exo2Ego settings. Rather than evaluating task performance, we aim to understand how PCCS forms consensus. We measure the L2 distance of different experts' prediction results and the labeled location of query masks, providing a quantitative assessment of alignment. Results indicate that individual experts exhibit distinct biases across scenarios, whereas PCCS aggregates these cues and consistently converges toward selections closer to the annotation space.

Visualization of HANDAL-X and DAVIS-17. Fig. 13 and Fig. 14 present qualitative results of our V^2-SAM method on the HANDAL-X and DAVIS-17 datasets, respectively. Across both indoor and outdoor scenes, V^2-SAM accurately localizes objects given only a single query frame, demonstrating strong consistency between query masks and predicted masks across subsequent viewpoints. On HANDAL-X, the model robustly segments small, elongated tools under large appearance variations and background clutter. On DAVIS-17, V^2-SAM generalizes to complex scenes involving articulated motion, occlusion, and diverse object categories, while maintaining precise object boundaries and temporal coherence. These visualizations validate that V^2-SAM effectively transfers query-driven segmentation cues across scenes and object instances.



(a) Ego2Exo. (b) Exo2Ego.

Figure 8. Qualitative comparison with Ref-SAM on the Ego-Exo4D dataset under two cross-view settings. The left column in each subfigure shows the query image, followed by predictions from Ref-SAM and our method. Our approach produces more accurate and consistent cross-view localization across both Ego2Exo and Exo2Ego scenarios.

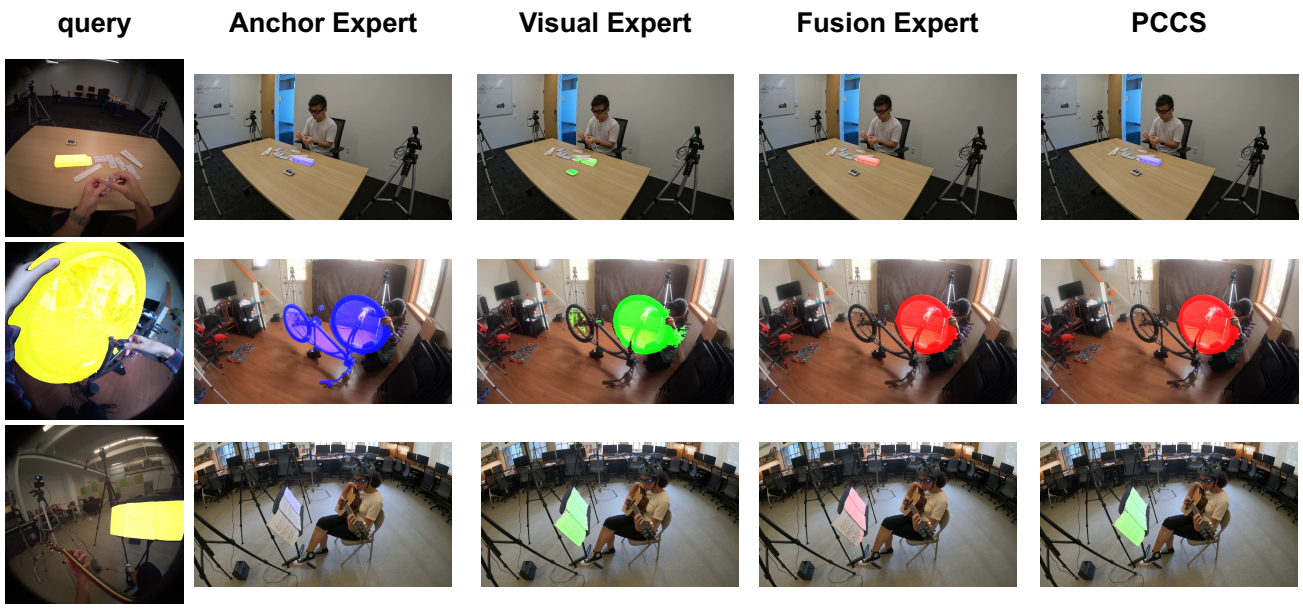


Figure 9. Comparison of selection results among individual experts and the PCCS on the Ego2Exo task. Each expert demonstrates varying strengths in interpreting specific first-person perspectives, while the PCCS leverages consensus across experts to improve overall selection consistency.

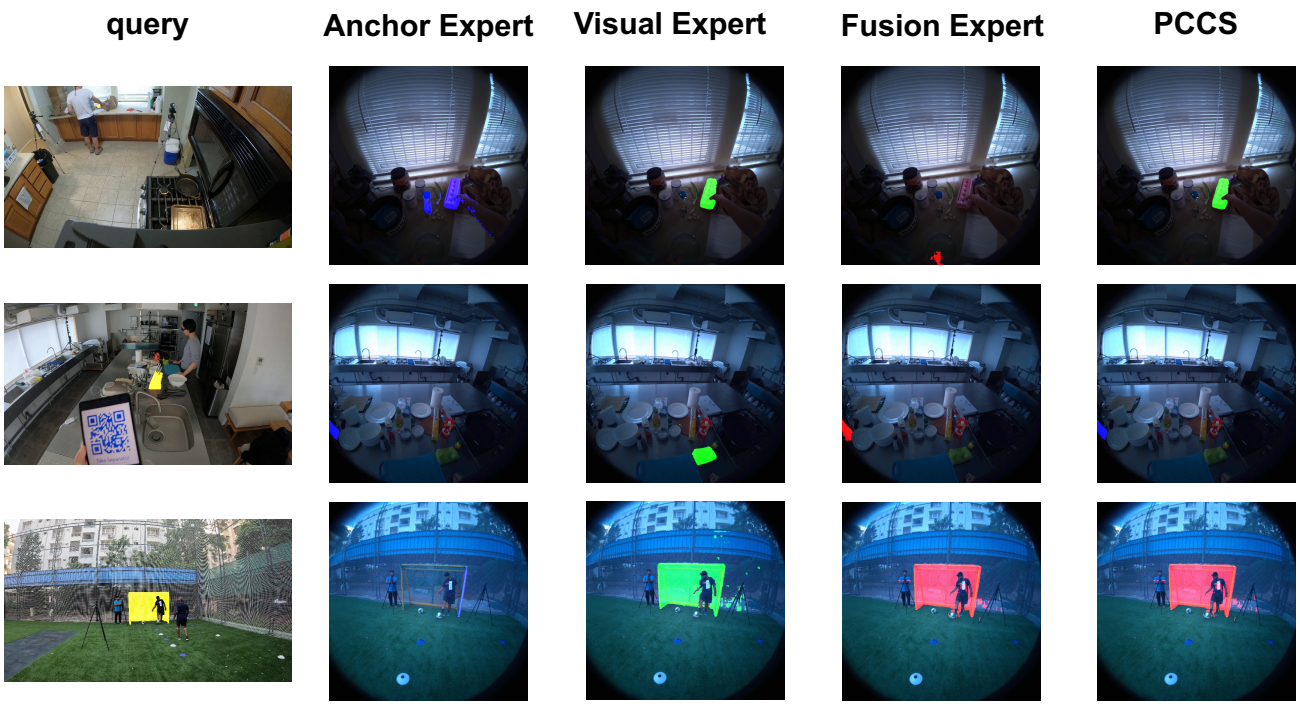


Figure 10. Comparison of selection results among individual experts and the PCCS on the Exo2Ego task. Experts show diverse interpretative preferences for third-person viewpoints, and the PCCS consolidates these judgments to yield more robust and balanced decisions.

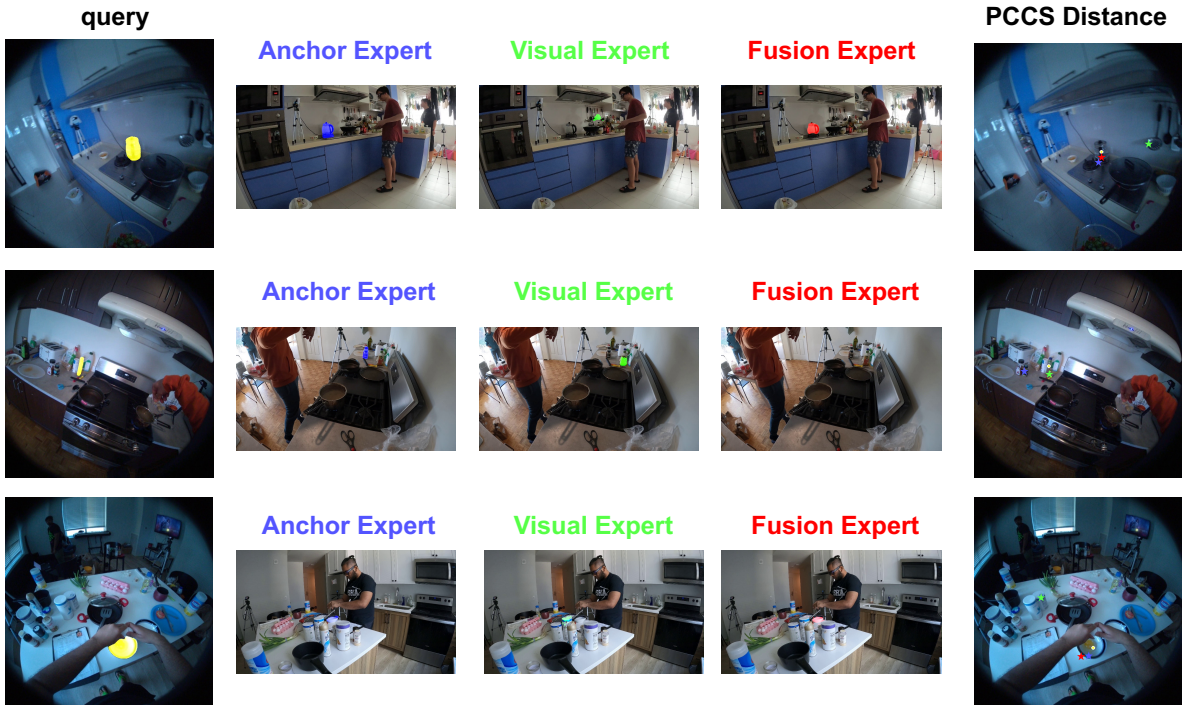


Figure 11. Ego2Exo Analysis. We quantify alignment by measuring the distance between the predicted locations of the Anchor Expert, Visual Expert, and Fusion Expert (colored accordingly) and the ground-truth query-mask annotations. Benefiting from integrating heterogeneous expert preferences, PCCS selects the expert whose prediction is closest to the annotation centroid.

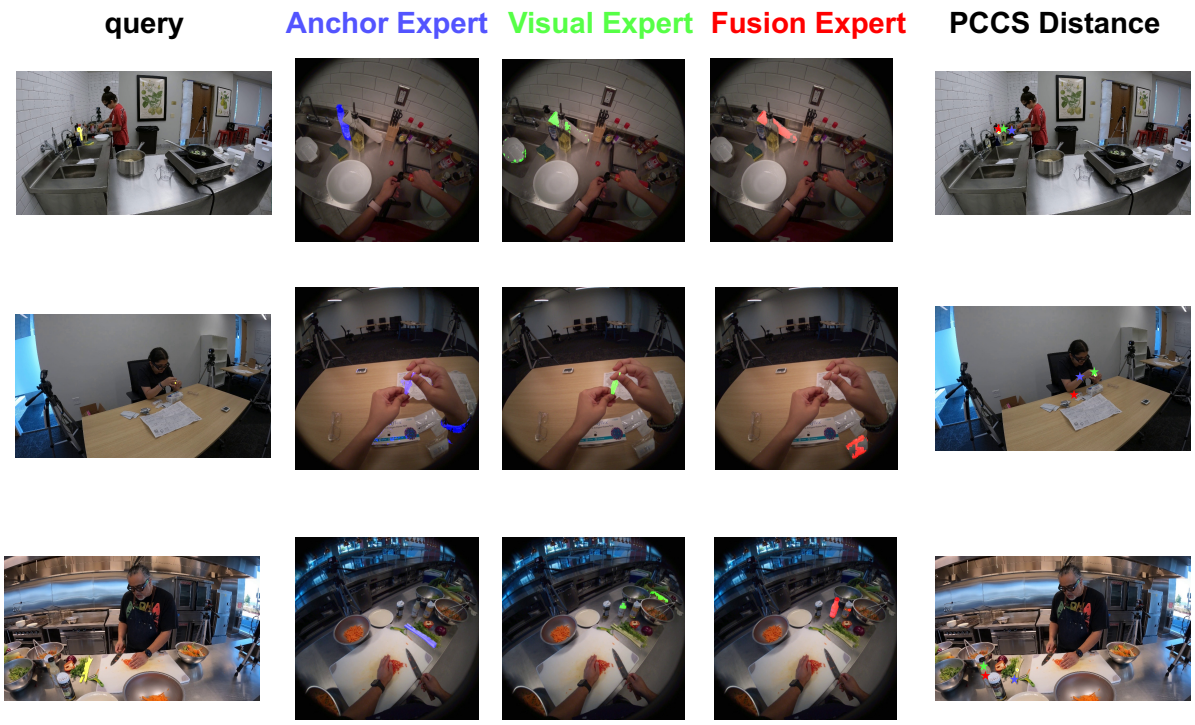


Figure 12. Exo2Ego Analysis. We quantify alignment by measuring the distance between the predicted locations of the Anchor Expert, Visual Expert, and Fusion Expert (colored accordingly) and the ground-truth query-mask annotations. Benefiting from integrating heterogeneous expert preferences, PCCS selects the expert whose prediction is closest to the annotation centroid.

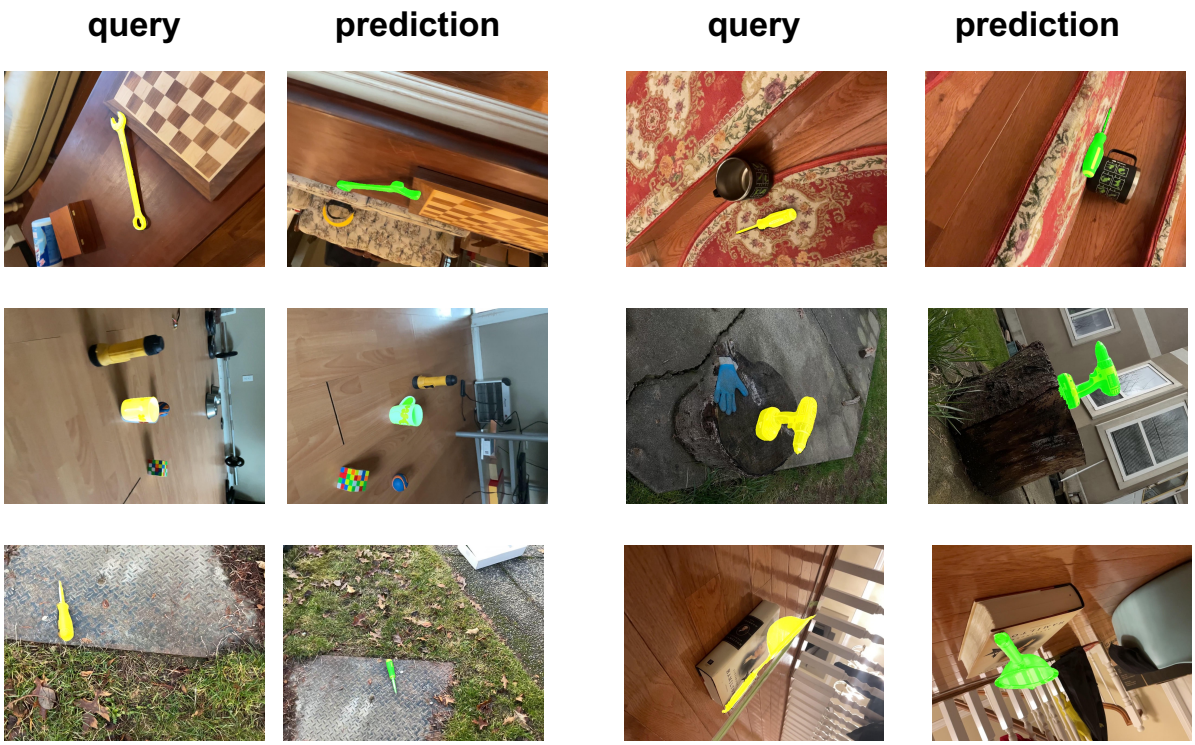


Figure 13. Visualization of the prediction results of our V^2-SAM method on the HANDAL-X dataset.

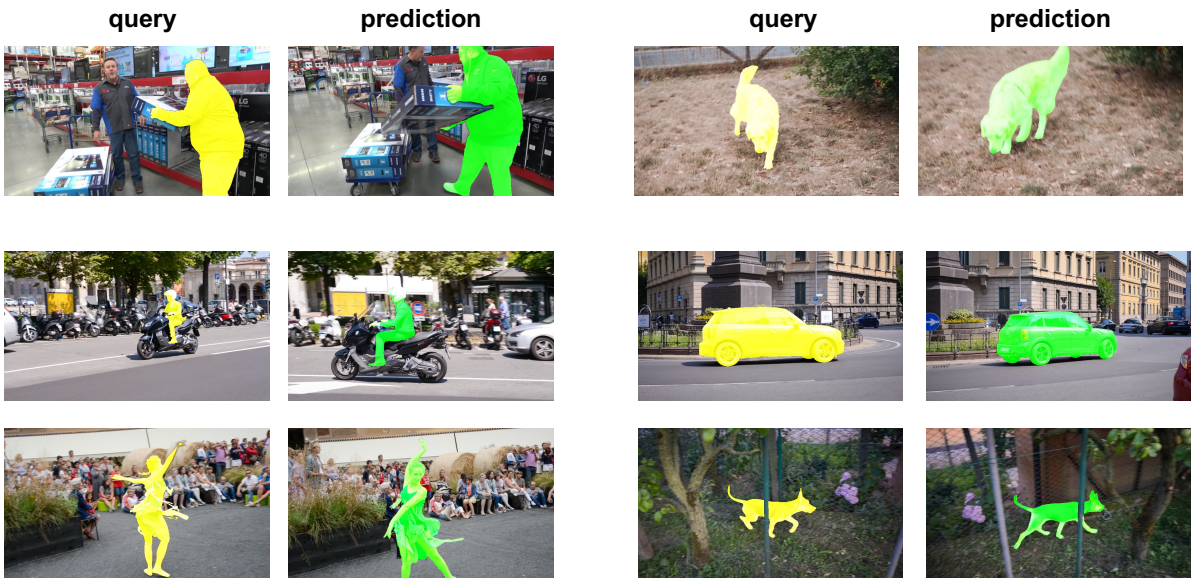


Figure 14. Visualization of the prediction results of our V^2-SAM method on the DAVIS-17 dataset.