Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Training-free open-vocabulary video instance segmentation via region-prior-guided multimodal reranking and vocabulary-driven re-segmentation

Gen Li
Guangdong Provincial Second Hospital of Traditional Chinese Medicine (Guangdong Provincial Engineering Technology Research Institute of Traditional Chinese Medicine)

Mingyong Zhang
Guangdong Provincial Second Hospital of Traditional Chinese Medicine (Guangdong Provincial Engineering Technology Research Institute of Traditional Chinese Medicine)

Wenfeng Zhang
809117128@qq.com

Guangdong Provincial Second Hospital of Traditional Chinese Medicine (Guangdong Provincial Engineering Technology Research Institute of Traditional Chinese Medicine)

Research Article

Keywords: Open-vocabulary video instance segmentation, Multimodal large language model, Region-aware CLIP, Option-order bias, Product of experts, Segment Anything

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11654739/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Training-free open-vocabulary video instance segmentation via region-prior-guided multimodal reranking and vocabulary-driven re-segmentation

Gen Li[1,2†], Mingyong Zhang[1,2†], Wenfeng Zhang[1,2*]

[1]Guangdong Provincial Second Hospital of Traditional Chinese Medicine (Guangdong Provincial Engineering Technology Research Institute of Traditional Chinese Medicine), Guangzhou, 510095, Guangdong, China.
[2]Guangdong Provincial Key Laboratory of Research and Development in Traditional Chinese Medicine, Guangzhou, 510095, Guangdong, China.

*Corresponding author(s). E-mail(s): 809117128@qq.com;
Contributing authors: 1148865278@qq.com;
†These authors contributed equally to this work.

Abstract

Open-vocabulary video instance segmentation (OV-VIS) asks a model to segment, track and name every object in a video, with names drawn from a vocabulary of more than a thousand fine-grained categories, most of which never appear in the training data of VIS models. We present CtxVIS, a training-free framework that separates localization from recognition and composes off-the-shelf foundation models. For every tracklet, a region-aware CLIP aggregated over eight temporally spread views supplies a prior over the full vocabulary and a shortlist of 15 candidates. A multimodal large language model (MLLM) then answers one multiple-choice question over the shortlist, and the two distributions are combined by a tempered log-linear product of experts. A diagnostic study shows that errors come mainly from reranking rather than from missing candidates, and that the MLLM prefers late option letters, which collides with options ordered by the prior; asking the question in forward and reverse order and averaging the two letter distributions removes this bias. Finally, the per-video category shortlist produced by CtxVIS prompts a concept-driven segmenter (SAM 3), whose detections are re-scored with CtxVIS evidence. On LV-VIS (1,196 categories), with every setting chosen on a 100-video development split and results reported on all 836 validation videos and on the 736 untouched ones, CtxVIS reaches 68.51 mask AP with ground-truth tracklets (+6.20 over Alpha-CLIP) and 41.89 AP fully

automatically, compared with 24.70 for OVFormer and 28.28 for OV-DVIS++ re-evaluated from their released weights, without any training.

Keywords: Open-vocabulary video instance segmentation, Multimodal large language model, Region-aware CLIP, Option-order bias, Product of experts, Segment Anything

1 Introduction

Video instance segmentation (VIS) jointly detects, segments, tracks and classifies object instances in a video [1, 2]. Classical VIS benchmarks contain tens of categories, and the strongest models are query-based video segmenters trained on these closed label sets [3–6]. Many applications that motivate VIS, such as video editing, augmented reality, robotics and content retrieval, cannot fix the label set in advance. Open-vocabulary VIS (OV-VIS) removes this restriction: at test time the model receives an arbitrary vocabulary and must name each tracked instance with one of its entries. The LV-VIS benchmark [7] makes the task concrete with 1,196 categories, 555 of which are absent from its training split, and large-vocabulary video benchmarks such as TAO and BURST [8, 9] pose the same problem.

Current OV-VIS models extend closed-set video segmenters with CLIP [10] text embeddings as the classifier and are trained on a limited set of base categories [7, 11–15]. Recognition remains the bottleneck: on LV-VIS val, the released OVFormer [13] and OV-DVIS++ [15] models reach 24.70 and 28.28 AP in our re-evaluation. In parallel, foundation models have made the two halves of the task individually much stronger. SAM 2 [16] produces and propagates class-agnostic masks through a video, SAM 3 [17] segments every instance of a concept given as a short noun phrase, region-aware CLIP models such as Alpha-CLIP [18] embed a masked region instead of the whole image, and multimodal large language models (MLLMs) [19–21] reason about fine-grained visual differences in natural language. Training-free compositions of such models are already competitive in open-vocabulary 3D instance segmentation [22, 23]. This raises the question we study: *can a training-free composition of foundation models outperform trained OV-VIS models on a thousand-category benchmark, and which design decisions actually matter?*

A direct transfer does not work. Following OV3D-CG [23], we let an MLLM describe the largest view of each ground-truth tracklet and pick a label from the full vocabulary; this yields only 24.06 AP on LV-VIS val, far below the 62.31 AP of a plain Alpha-CLIP retrieval on the same masks. MLLMs are known to be weak at classification over large label sets [24], but they are good at choosing among a few plausible candidates. We therefore use a region prior to restrict the choice, ask the MLLM a single multiple-choice question over the shortlist, and fuse its answer with the prior. Making this simple idea work well required three findings that we document in detail. (i) The MLLM's letter distribution is badly over-confident, so a tempered product with the prior that keeps the prior dominant works far better than trusting the MLLM. (ii) The model has a strong preference for late option letters, which interacts with the natural choice of listing candidates in prior order; presenting the options

in both orders and averaging the answers cancels this bias. (iii) Once recognition is strong, the class-agnostic tracklets limit the fully automatic pipeline, and the shortlist itself is the best prompt for a concept-driven segmenter.

Our contributions are:

- A training-free OV-VIS framework, CtxVIS, in which a multi-view region prior shortlists candidates, an MLLM reranks them with one multiple-choice query, and the two are combined by a tempered log-linear product of experts. With ground-truth masks it improves Alpha-CLIP from 62.20 to 67.88 AP (top-1 accuracy +8.34 points, McNemar $p = 1.3 \times 10^{-38}$) on held-out videos.
- A diagnostic analysis of where errors come from: candidate recall versus reranking, calibration of both experts, the effect of temporal view aggregation inside and outside the pipeline, and the option-position bias of the MLLM. The analysis motivates a bidirectional option order (CtxVIS-BO) that removes the bias at the cost of one additional query.
- A vocabulary-driven re-segmentation stage that turns the per-video CtxVIS shortlist into SAM 3 concept prompts and re-scores the detections with CtxVIS evidence. The fully automatic pipeline reaches 41.89 AP on LV-VIS val, 13.61 AP above the strongest trained model we re-evaluated.
- A strict protocol: all choices are made on a 100-video development split, results are reported on the remaining 736 videos as well as on the full validation set, and differences are tested with paired exact McNemar tests and video-level bootstrap intervals. Baselines are re-run from released weights with the same evaluator.

2 Related work

Video instance segmentation.

VIS was introduced with YouTube-VIS [1] and extended to heavy occlusion [2] and large vocabularies [8, 9]. Query-based segmenters [3] dominate the field: MinVIS [4] tracks image-level queries without video training, IDOL [5] uses contrastive association for online inference, and DVIS [6] and DVIS++ [15] decouple segmentation, tracking and refinement. Video object segmentation, which tracks given objects without naming them, has seen a parallel line of work on memory networks, motion cues and efficient mask transformers [25–27]. All of these models are trained on closed label sets.

Open-vocabulary VIS.

LV-VIS [7] defined OV-VIS with a large-vocabulary benchmark and the OV2Seg baseline, which replaces the classifier of a video segmenter with CLIP text embeddings. OpenVIS [11] adds an open-world proposal network and instance-guided CLIP features, BriVIS [12] models instance trajectories as Brownian bridges aligned with class texts, CLIP-VIS [14] adapts a frozen CLIP with class-agnostic mask generation and weighted classification, OVFormer [13] aligns query embeddings with the CLIP space, and OV-DVIS++ [15] trains the DVIS++ framework with a frozen CLIP backbone. These methods learn the segmenter and the vision–language alignment from base categories, so recognition of novel and rare categories relies on the CLIP embedding of a mask-pooled feature. CtxVIS instead keeps every model frozen and invests the effort

in recognition with an MLLM, which reasons about a short list of candidates rather than embedding a feature.

Open-vocabulary detection and segmentation in images.

Image-level open-vocabulary detectors distil or align CLIP features with region proposals [28–32], and open-vocabulary segmenters classify mask proposals with CLIP [33–35]. Training-free pipelines that chain a detector, SAM and CLIP [36, 37] or lift 2D foundation-model features to 3D [22, 23, 38] show that composition can rival training. OV3D-CG [23] is closest in spirit: an MLLM describes the best view of a 3D instance and the description is matched to the vocabulary. We show that this description-first design transfers poorly to a thousand-category video vocabulary and that the MLLM is more useful as a reranker of a region prior.

Region-aware recognition and visual prompts.

CLIP features of a crop include the background, and mask pooling loses context. Alpha-CLIP [18] feeds the mask as an extra alpha channel, and visual prompts such as red circles [39] or numbered marks [40] steer attention to a region. Region-level MLLMs [41, 42] are trained to refer to regions. We use Alpha-CLIP for the prior and compare crops with mask overlays as the visual prompt for the MLLM; LLM-generated class descriptions [43] are evaluated as an option-definition variant.

MLLMs as classifiers and option-order bias.

Zhang et al. [24] showed that visually grounded language models classify worse than CLIP over large label sets and attributed this to data rather than architecture. When an LLM answers multiple-choice questions, its answer depends on the order and the letters of the options [44, 45]; similar position effects appear in pairwise judgement [46], in-context learning (recency bias) [47] and long-context retrieval [48]. Proposed remedies permute the options and average the answers, or estimate and remove a prior over letters [44, 47]. We measure the bias of Qwen2.5-VL in a visual recognition setting where the option order is not random but follows the prior ranking, which makes the bias systematically harmful, and show that a single reversed query suffices to cancel it.

Combining experts.

Combining calibrated probabilistic classifiers by products or sums is classical [49, 50]; tempering the factors controls their relative confidence, and temperature scaling is the standard post-hoc calibration [51]. Our fusion is a tempered product of two experts whose two parameters are chosen on the development split.

Promptable segmentation and tracking.

SAM [52] segments from points, boxes or masks, and SAM 2 [16] extends it to videos with a streaming memory. DEVA [53] decouples image-level segmentation from temporal propagation, and Grounded-SAM [36] prompts SAM with open-set detections. SAM 3 [17] segments and tracks all instances of a concept given as a noun phrase.

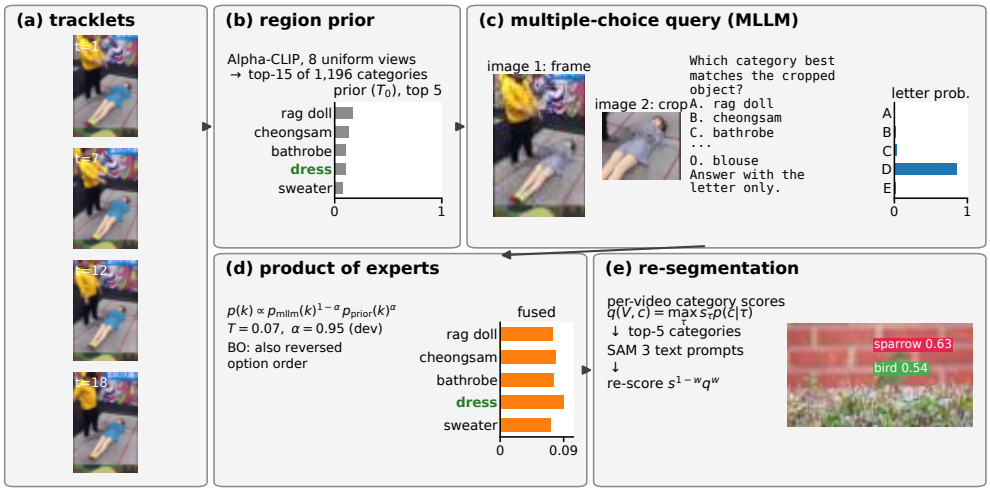


Fig. 1 Overview of CtxVIS. (a) Class-agnostic tracklets. (b) Alpha-CLIP features of eight uniformly spread views are averaged into a region prior over the full vocabulary, and the top-15 categories form the shortlist. (c) The MLLM sees the best-quality frame and a close-up crop and answers one multiple-choice question over the shortlist; CtxVIS-BO additionally asks with the options in reverse order. (d) The letter distribution and the retempered prior are combined by a log-linear product of experts. (e) The per-video category scores select a five-category shortlist that prompts SAM 3, and the SAM 3 detections are re-scored with CtxVIS evidence. Panels (a)–(d) follow one held-out track; (b) shows the prior at CLIP's temperature T_0, (d) the fused distribution at the retempered prior, and (e) the output on another held-out video.

Prompting SAM 3 with the full LV-VIS vocabulary would require more than a thousand prompts per video; CtxVIS provides the small per-video vocabulary that makes concept prompting practical.

3 Method

Figure 1 gives an overview. Given a video $V = \{I_t\}_{t=1}^{T}$ and a vocabulary $\mathcal{C}$, CtxVIS first obtains class-agnostic tracklets $\tau = (\{m^t\}_{t=1}^{T}, s_\tau)$ with per-frame masks m^t (possibly empty) and a localization score s_τ (Sect. 3.6). Each tracklet is recognized by a region prior (Sect. 3.2) and a shortlist multiple-choice query to an MLLM (Sect. 3.3), fused by a tempered product of experts (Sect. 3.4). The resulting category scores of a video then drive concept-prompted re-segmentation (Sect. 3.5). No component is trained or fine-tuned.

3.1 Problem formulation

The output for a video is a set of scored, labelled mask tubes $\{(\tau, c, \sigma)\}$ evaluated by video mask AP [1]. Like most open-vocabulary segmenters we factorize the task into class-agnostic localization and recognition. Recognition assigns every tracklet a distribution $p(c \mid \tau)$ over a shortlist of $\mathcal{C}$; each tracklet contributes its three most probable categories with score $\sigma = s_\tau\, p_{\text{obj}}(\tau)\, p(c \mid \tau)$, where p_{obj} is an objectness term

described in Sect. 3.3. Emitting several hypotheses per tracklet is standard for AP, which rewards a correct label even at a lower rank.

3.2 Multi-view region prior

Let $\mathcal{V}(\tau) = \{t : m^t \neq \emptyset\}$ be the frames in which the tracklet is visible. We take $N = 8$ frames spread uniformly over $\mathcal{V}(\tau)$ in time. For each view we cut a square window around the mask with side $1.5\max(w, h)$ of its bounding box, zero-pad it where it leaves the image, and encode the window with the Alpha-CLIP image encoder using the mask as the alpha channel, giving unit-norm embeddings $f_1, \ldots, f_N$. The track embedding is their normalized mean $z = \bar{f}/\|\bar{f}\|$. Each category c is represented by the normalized mean of the Alpha-CLIP text embeddings of six prompt templates (e.g. "a photo of a {c}.", "a video frame showing a {c}."), w_c. The cosine similarities $s_c = z^\top w_c$ define the prior

$$p_0(c \mid \tau) = \frac{\exp(s_c/T_0)}{\sum_{c' \in \mathcal{C}} \exp(s_{c'}/T_0)}, \qquad T_0 = 0.01, \tag{1}$$

where T_0 is CLIP's logit scale. The $K = 15$ most probable categories form the shortlist $\mathcal{K}(\tau) = (k_1, \ldots, k_K)$, ordered by decreasing prior. Within the shortlist the prior is retempered to $p_{\text{prior}}(k) \propto \exp(s_k/T)$, with T chosen on the development split (Sect. 3.4).

Averaging views is the only temporal reasoning in CtxVIS. We also evaluated quality-weighted means, trimmed means, log-sum-exp pooling of per-view probabilities, and temporal storyboards shown to the MLLM (Sect. 5.5).

3.3 Shortlist multiple-choice query

Visual input.

The MLLM sees two images: the full frame I_{t^*} for context, and a close-up crop around the object in the same frame (bounding box enlarged by 1.5). The frame t^* is chosen for visibility, completeness and sharpness. Among the five frames with the largest value of $a_t(1 - 0.2\, b_t)$, where a_t is the mask area and $b_t \in \{0, \ldots, 4\}$ the number of image borders the box touches, we take

$$t^* = \arg\max_t \; \frac{a_t}{a_{\max}} \, (1 - 0.2\, b_t) \left(\frac{1}{2} + \frac{1}{2} \frac{\lambda_t}{\lambda_{\max}} \right), \tag{2}$$

where λ_t is the variance of the Laplacian inside the box (a sharpness measure) and $a_{\max}$ and $\lambda_{\max}$ are the maxima over the candidates. Frames are resized to at most 768 pixels on the longer side.

Question.

The prompt is

```
Image 1 is a frame from a video. Image 2 is a close-up crop of one object
from that frame. Which category best matches the cropped object?
```

`A.` $\{k_1\}$ `B.` $\{k_2\}$ `...` `O.` $\{k_{15}\}$
`Answer with the letter of the best option only.`

with category names written as plain words. Instead of generating text, we run a single forward pass and read the logits ℓ_j of the option letters at the first answer position. Their softmax gives the MLLM distribution over the shortlist,

$$p_{\text{mllm}}(k_j) = \frac{\exp \ell_j}{\sum_{i=1}^{K} \exp \ell_i}. \tag{3}$$

This costs one prefill of two images and about two hundred text tokens, needs no output parsing, and returns a full distribution that can be fused.

Rejecting spurious tracklets.

Automatically generated tracklets include background regions and object parts. For them we append the option "none of the above (background, part of an object, or not a distinct object)". Its probability gives $p_{\text{obj}} = 1 - p_{\text{mllm}}(\text{reject})$, and the remaining letters are renormalized. With ground-truth tracklets the option is omitted and $p_{\text{obj}} = 1$.

Bidirectional option order (CtxVIS-BO).

Listing the shortlist in prior order is natural, but it places the most likely candidates in the first letters. Sect. 5.3 shows that Qwen2.5-VL prefers late letters, so the prior order systematically works against the best candidates. CtxVIS-BO asks the same question a second time with the options in reverse order, maps both letter distributions back to the candidates, and takes their normalized geometric mean

$$p^{\text{BO}}_{\text{mllm}}(k) \propto \sqrt{p^{\rightarrow}_{\text{mllm}}(k)\, p^{\leftarrow}_{\text{mllm}}(k)}. \tag{4}$$

Reversal moves every candidate from position j to $K+1-j$, so a monotonic position preference is cancelled to first order, which a random permutation does not guarantee.

3.4 Tempered product of experts

The MLLM is accurate when the decision depends on fine visual attributes that it can name, while the prior is better calibrated and covers categories that the MLLM does not recognize. We combine them by a tempered product of experts [49, 50]

$$p(k \mid \tau) \propto p_{\text{mllm}}(k)^{1-\alpha}\, p_{\text{prior}}(k)^{\alpha}, \qquad p_{\text{prior}}(k) \propto \exp(s_k/T), \tag{5}$$

which is equivalent to adding the MLLM log-probabilities, scaled by $(1-\alpha)$, to the CLIP similarities scaled by α/T. Two numbers therefore control the fusion: the prior temperature T, which sets how peaked the prior is, and the weight α. We select them by grid search on the development split ($T \in [0.01, 0.2]$, $\alpha \in [0, 1]$) and obtain $T = 0.07$ and $\alpha = 0.95$, an interior optimum (Fig. 4). Despite the large α, the MLLM changes the decision often: its letter distribution is very peaked, so even the exponent 0.05

moves the ranking between candidates that the prior finds close. We also compared linear mixing and gates that predict a per-tracklet α from the confidence of both experts (Table 3).

3.5 Vocabulary-driven re-segmentation

With ground-truth tracklets the recognition stage is evaluated in isolation. In the fully automatic setting the class-agnostic tracklets become the limiting factor: they fragment objects, cover parts, and miss small instances. Concept-prompted segmenters such as SAM 3 find all instances of a named concept, but a thousand prompts per video are impractical. CtxVIS supplies the missing piece, a per-video shortlist of the categories that are likely present.

First, CtxVIS recognizes the SAM 2 tracklets of the video (Sect. 3.6) and aggregates video-level category scores

$$q(V, c) = \max_{\tau \in V} \; s_\tau \, p_{\text{obj}}(\tau) \, p(c \mid \tau). \tag{6}$$

The five categories with the highest q become text prompts for the SAM 3 video predictor, which detects every instance of each concept on the first frame, tracks it through the video, and scores a detection d by its mean probability s_d over the tracked frames. The SAM 3 score measures how well the mask matches the phrase, but not whether the phrase is the right category for the video. We therefore re-score each detection of concept c with the CtxVIS evidence for that concept,

$$\sigma_d = s_d^{1-w} \, q(V, c)^{\,w}, \tag{7}$$

with $w = 0.25$ chosen on the development split. The re-scored SAM 3 detections form the final output.

3.6 Class-agnostic tracklets

The SAM 2 tracklets are produced without any category information. On keyframes every 10 frames, the SAM 2.1 Hiera-L automatic mask generator samples a 32×32 point grid; masks with predicted IoU above 0.8, stability above 0.9 and an area between 0.15% and 50% of the image are kept, sorted by predicted IoU $\times$ stability, and de-duplicated (IoU < 0.7, at most 30 per keyframe). Masks from the first keyframe initialize objects; masks from later keyframes are added only if less than 70% of their area is already covered by a tracked object and their IoU with every tracked mask is below 0.5. All objects (at most 50) are propagated forward and backward with the SAM 2 video predictor. A tracklet's score is the mask quality of its seed multiplied by the square root of the fraction of frames in which it is visible, and tracklets with a spatio-temporal IoU above 0.6 are merged.

4 Experimental setup

Dataset and protocol.

We evaluate on LV-VIS [7], whose validation set has 836 annotated videos with 3,719 instance tracks (22.9 frames per video on average) over 1,196 categories (641 base, 555 novel). The test annotations are not public, and the related BURST and TAO benchmarks [8, 9] have non-exhaustive labels, so we use LV-VIS val throughout and guard against overfitting by splitting it. A fixed random subset of 100 videos (444 tracks) is the *development* split on which every design choice and every hyper-parameter is made. We report results on the full validation set (*val*, comparable with the literature) and on the remaining 736 videos (3,275 tracks), which we call *held-out*. A further random subset of 50 held-out videos is used only for timing.

Settings and metrics.

In the *oracle* setting the ground-truth tracks are given without labels and every method must name them. This isolates recognition, as in OV3D-CG [23]. In the *fully automatic* setting a method produces its own tracklets. We report video mask AP, AP_{50}, AP_{75} and AP on base (AP_b) and novel (AP_n) categories, computed by our re-implementation of the official LV-VIS evaluator, which we verified against the official script on exported results. In the oracle setting we also report per-track top-1 accuracy. Note that base categories in LV-VIS are frequent and fine-grained (many kinds of vehicles, food and tools) while many novel categories are visually distinctive, so top-1 accuracy is lower on base categories for every method. Differences in top-1 accuracy between two methods are tested with the exact two-sided McNemar test on the per-track outcomes [54], and 95% confidence intervals are obtained by resampling videos 2,000 times [55].

Models.

The region prior uses Alpha-CLIP ViT-L/14 (GRIT-20M, full fine-tuning) with the OpenAI CLIP text encoder. The MLLM is Qwen2.5-VL-7B-Instruct [20] with at most 401,408 pixels per image; Qwen3-VL-8B [56], InternVL3-8B [21] and Qwen2.5-VL-3B are evaluated as alternatives. Tracklets come from SAM 2.1 Hiera-L [16], and re-segmentation uses the released SAM 3 model [17] with a detection threshold of 0.5. All models run in bf16 on one NVIDIA RTX 4090D (24 GB). Decoding is greedy and every result is deterministic.

Baselines.

In the oracle setting we compare with (i) CLIP ViT-L/14 on context crops of the three largest views [10, 22]; (ii) Alpha-CLIP with the mask as alpha channel on the three largest views [18] and (iii) on eight uniform views (our prior alone); (iv) an OV3D-CG-style pipeline [23] in which Qwen2.5-VL describes the largest view, names the object given the full label list, and the name is matched to the vocabulary with CLIP; and (v) the MLLM alone on our shortlist ($\alpha = 0$). In the fully automatic setting we evaluate the released OVFormer (ResNet-50 and Swin-B, trained on the LV-VIS

Table 1 Recognition of ground-truth LV-VIS tracklets (oracle setting). AP and top-1 accuracy (%) on the full validation set (val, 836 videos) and on the held-out videos not used for any design decision (736 videos). Δtop-1 is the paired difference to Alpha-CLIP (3 views) on held-out tracks, with McNemar p. Calls: MLLM forward passes per tracklet.

		val		held-out				vs. Alpha-CLIP	
Method	Calls	AP	top-1	AP	AP_b	AP_n	top-1	Δtop-1	p
CLIP ViT-L/14, 3 views	0	48.87	33.10	48.46	34.45	58.84	32.43	−9.16	1.2×10^{-36}
Alpha-CLIP, 3 largest views	0	62.31	42.24	62.20	48.61	72.26	41.59	–	–
Alpha-CLIP, 8 views (prior)	0	62.63	42.67	61.98	47.24	72.90	41.65	+0.06	0.94
OV3D-CG-style	2	24.06	21.30	23.81	16.27	29.40	21.47	−20.12	4.9×10^{-99}
MLLM only on shortlist	1	59.98	47.51	59.31	47.70	67.92	47.39	+5.80	5.2×10^{-11}
CtxVIS, two-step	2	68.03	49.42	67.66	52.41	78.95	49.37	+7.79	4.6×10^{-28}
CtxVIS	1	68.51	50.39	67.88	53.22	78.75	49.92	+8.34	1.3×10^{-38}
CtxVIS-BO	2	**69.17**	**51.22**	**68.40**	**53.89**	**79.15**	**50.72**	+9.13	7.0×10^{-47}

training split) [13] and OV-DVIS++ (online ResNet-50, online and offline ConvNeXt-L, trained on COCO) [15] on LV-VIS val with all 1,196 categories. Our OVFormer numbers (22.11 and 24.70 AP) agree with the 21.9 and 24.7 AP reported by its authors, which validates the evaluator. We report only numbers that we measured ourselves under one protocol; other published methods are discussed in the text.

Use of AI tools.

Apart from the MLLMs that are part of the method, an AI assistant was used in a supporting role: to help write the evaluation and plotting scripts and to draft and language-edit parts of the text. All methods, experimental results and conclusions were checked by the authors against the experiment records, and the authors take full responsibility for the content of this article.

5 Results

5.1 Recognition with ground-truth tracklets

Table 1 compares recognition methods on ground-truth tracklets. Region-aware features matter: Alpha-CLIP improves on mask-cropped CLIP by 13.74 AP on held-out videos. The OV3D-CG-style transfer, which asks the MLLM for a free-form description and a label from the full vocabulary, reaches only 23.81 AP, and the MLLM alone on our shortlist (59.31 AP) is also below Alpha-CLIP, although its top-1 accuracy (47.39%) is higher. The MLLM is a good ranker but a poorly calibrated one, which hurts AP because AP ranks all predictions of a category across videos by their scores. Fusing the two experts gives CtxVIS 68.51 AP on val and 67.88 AP on held-out videos, +5.68 AP over Alpha-CLIP with the same masks. The per-track gain in top-1 accuracy is +8.34 points (95% CI [6.82, 9.90]; 370 tracks improve, 97 get worse; $p = 1.3 \times 10^{-38}$). The bidirectional query adds a further +0.79 points of top-1 accuracy over CtxVIS (59 vs. 33 tracks, $p = 0.009$) and reaches 68.40 AP.

A two-step describe-then-choose variant, in which the MLLM first writes a short description of the object and then answers the same question, performs on par (67.66 AP) at twice the cost, so CtxVIS uses the single call. Novel categories gain more

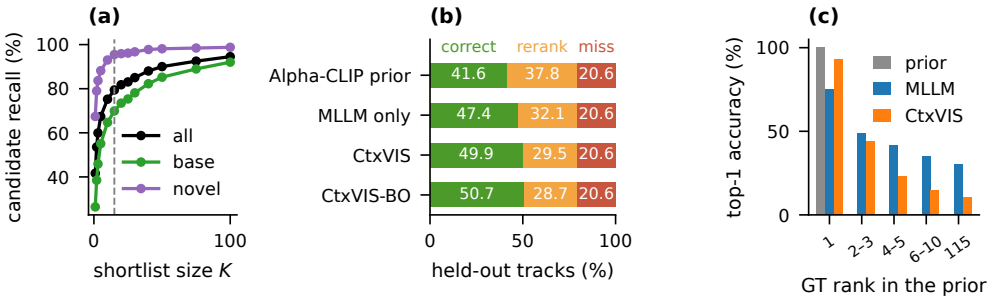


Fig. 2 Error decomposition on held-out tracks (oracle setting). (a) Candidate recall of the Alpha-CLIP prior as a function of the shortlist size K; the dashed line marks $K = 15$. (b) Outcome of each method's top-1 prediction: correct, ground truth in the shortlist but ranked below another candidate (reranking error), or ground truth outside the shortlist (recall error). (c) Top-1 accuracy of CtxVIS as a function of the rank of the ground truth in the prior.

AP than base categories (AP_n 78.75 vs. 72.26 for Alpha-CLIP), consistent with the MLLM recognizing distinctive objects by name that a CLIP embedding confuses with visually similar ones.

5.2 Where do the errors come from?

Recognition in the oracle setting can fail in two ways: the correct category is missing from the shortlist, or it is present but another candidate is chosen. Figure 2 separates the two. With $K = 15$ the shortlist contains the ground truth for 79.4% of the held-out tracks, and recall grows slowly with K (83.1% at $K = 25$). Of all held-out tracks, 20.6% are lost because the ground truth is not in the shortlist, and 29.5% because CtxVIS ranks another candidate first. Reranking is thus the larger source of error. A perfect reranker of the same shortlists would reach 90.01 AP, far above the current 67.88, whereas enlarging the shortlist to $K = 25$ does not help (67.87 AP, Table 4): the extra candidates add recall but also distractors. The gain of CtxVIS over the prior comes from tracks whose ground truth the prior ranks second to fifth (Fig. 2c), exactly where the MLLM is asked to choose between visually similar categories. The price is that CtxVIS loses 7.3% of the tracks that the prior already ranks correctly; overall 370 tracks are corrected and 99 broken.

5.3 Option-position bias of the MLLM

To test whether the answer depends on where a candidate appears, we repeated the held-out query with three other option orders while keeping the shortlist, the images and the prior fixed: the reverse of the prior order and two random permutations. Figure 3 shows a clear preference for late options. Under random permutations the first two letters are chosen in only 4.95% and 5.45% of the cases, against a uniform rate of 6.7%, and the MLLM alone is correct for 53.2% of the tracks whose ground truth is at positions A–C but for 64.0% when it is at positions D–O. In prior order, the prior's best candidate sits in position A; the MLLM picks it in 45.28% of the cases, compared with 50.23% when the same candidate is presented last. The bias is

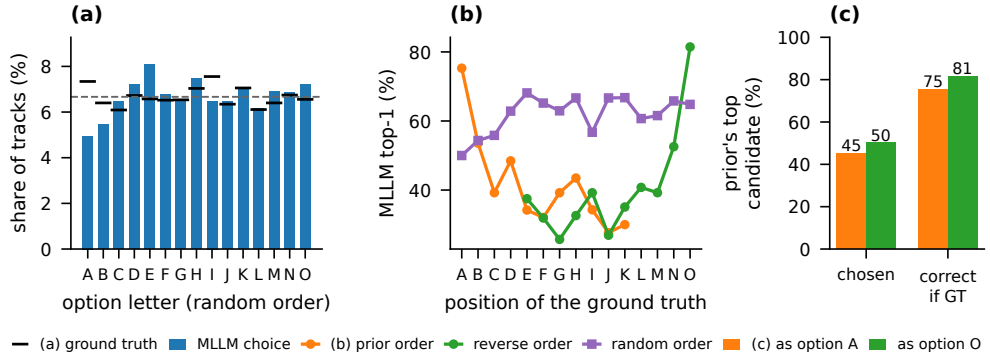


Fig. 3 Position bias of Qwen2.5-VL-7B on the 15-way recognition question (held-out tracks). (a) How often each option letter is chosen when the options are randomly permuted (bars), how often the ground truth is at that letter (black marks), and the uniform rate 1/15 (dashed). (b) Accuracy of the MLLM alone as a function of the position of the ground truth, for options in prior order, in reverse prior order, and in random order. (c) When the options follow the prior order, the prior's best candidate is option A; with reversed order it is option O.

not merely a nuisance: it pushes the MLLM away from the candidates that the prior considers most likely, in exactly the cases where the prior is right.

Table 2 quantifies how option order and combinations of orders affect the fused result. After fusion, single orders differ by less than one AP point, and the prior order is the weakest of them in AP. Combining two queries helps when they disagree: the forward and reverse queries choose the same option for 77.0% of the tracks, and at least one of them is correct for 53.9%, whereas a single query is correct for 47.36%. Their geometric mean (CtxVIS-BO) is the only two-query combination whose top-1 gain is significant, on held-out videos ($p = 0.009$) as well as on the full validation set ($p = 0.003$). In AP, the combinations lie within 0.5 points of each other. A second view instead of a second order, i.e. the crop of the largest-area frame, is much less complementary: for 31.9% of the tracks the largest frame is also the best-quality frame and the query is identical, and the two choices agree for 87.8%. Combining all four orders does not raise top-1 accuracy beyond CtxVIS-BO. Several combinations were evaluated on the same held-out data, so we regard the improvement of CtxVIS-BO as modest. It is, however, the combination predicted by the position analysis, and it is positive on both evaluation sets. The fully automatic pipeline uses the single query.

5.4 Fusion and calibration

Figure 4c explains why a strong prior weight is needed. The MLLM is over-confident: its mean top-1 confidence is 0.74 and its expected calibration error (ECE) is 0.27, whereas the prior at CLIP's own temperature T_0 is well calibrated (ECE 0.07). Retempering to $T = 0.07$ is not meant to calibrate: it deliberately flattens the prior over the shortlist (mean confidence 0.10), so that the sharply peaked MLLM factor can reorder candidates that the prior finds close, even with the exponent $1 - \alpha = 0.05$. In absolute terms the fused distribution is therefore under-confident (ECE 0.37), but its confidence increases monotonically with accuracy (0.09 → 24%, 0.14 → 55%, 0.22 →

Table 2 Option order and query combinations (held-out, oracle setting). All rows use the same shortlists, images and fixed fusion ($T = 0.07$, $\alpha = 0.95$); combinations average the letter distributions geometrically. Δtop-1 and p are paired against the single query in prior order (CtxVIS).

Queries	Calls	AP	top-1	Δtop-1	p
prior order (CtxVIS)	1	67.88	49.92	–	–
reverse prior order	1	68.30	50.29	+0.37	0.40
random order 1	1	68.49	49.37	−0.55	0.21
random order 2	1	68.45	49.56	−0.37	0.40
largest-area frame, prior order	1	68.02	49.59	−0.34	0.30
prior + reverse (CtxVIS-BO)	2	68.40	**50.72**	+0.79	0.009
prior + random 1	2	68.58	50.23	+0.31	0.33
random 1 + random 2	2	68.34	50.02	+0.09	0.88
best + largest-area frame	2	68.34	50.17	+0.24	0.29
random 1 + largest-area frame	2	**68.88**	50.26	+0.34	0.34
all four orders	4	68.74	50.32	+0.40	0.27

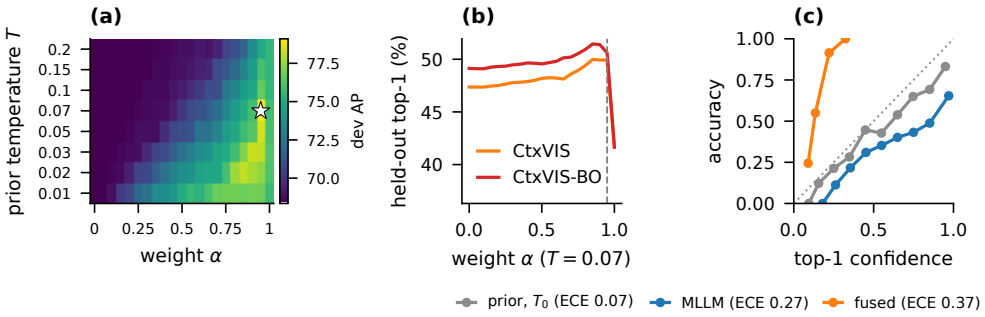


Fig. 4 Fusion of the two experts. (a) Development-split AP of the log-linear fusion over prior temperature T and weight α (describe-then-choose records, on which the setting was selected); the star marks the selected setting. (b) Held-out top-1 accuracy of CtxVIS and CtxVIS-BO as a function of α at $T = 0.07$; $\alpha = 0$ is the MLLM alone and $\alpha = 1$ the prior alone. (c) Reliability diagram of the top-1 confidence on held-out tracks for the shortlist prior at CLIP's temperature T_0, the MLLM and the fused distribution, with expected calibration error (ECE).

91% in the populated bins). This is what matters for AP, which depends only on the order of the scores across tracklets and is invariant to any monotonic transformation of them. The selected setting is an interior optimum on the development split, and accuracy varies smoothly around it (Fig. 4b). Table 3 compares fusion forms, each with its parameters selected on the development split: the log-linear product is better than linear mixing, and gates that predict a per-track weight from confidence features do not generalize from 100 development videos.

Table 3 Fusion forms, parameters selected on the development split (dev AP) and evaluated on held-out tracks (oracle setting, describe-then-choose records).

Fusion	Selected setting	dev AP	held-out AP	top-1
prior only ($\alpha = 1$)	–	75.69	61.19	41.65
MLLM only ($\alpha = 0$)	–	67.24	56.11	44.73
linear	$T = 0.03$, $\alpha = 0.85$	78.03	65.92	47.76
log-linear (ours)	$T = 0.07$, $\alpha = 0.95$	**79.10**	**67.66**	**49.37**
gated linear	$T = 0.03$, 6-feature gate	74.48	63.02	44.98
gated log-linear	$T = 0.07$, 6-feature gate	74.11	61.79	45.74

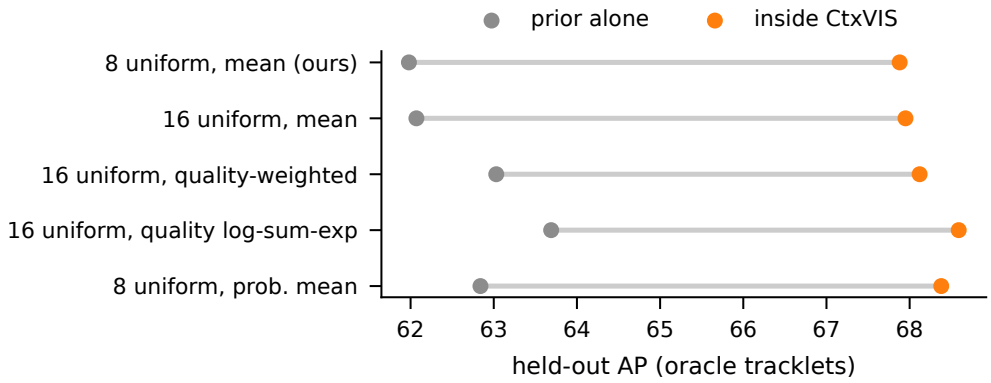


Fig. 5 Temporal aggregation of the region prior on held-out tracks. Grey: AP of the prior alone; colour: AP of the full CtxVIS when only the prior is replaced (shortlist and MLLM answers fixed). Gains of the prior largely vanish inside the pipeline.

5.5 Temporal aggregation: why more views stop helping

A video offers many views of an object, and aggregating them improves the prior substantially (Fig. 5). Alpha-CLIP on a single view gives 55.73 AP; the mean of eight uniformly spread views gives 61.98, and better pooling of sixteen views reaches up to 63.69 AP (quality-weighted log-sum-exp of per-view probabilities). The eight uniform views were selected on the development split, where they were 2.6 AP better than the three largest views; on held-out videos the two are equivalent (61.98 vs. 62.20 AP, top-1 $p = 0.94$). However, when we keep the shortlists and the MLLM answers fixed and only exchange the prior inside the fusion, the differences shrink to at most 0.71 AP and none of them is significant in top-1 accuracy ($p > 0.31$). The improvements of better pooling occur mostly in the order of the top few candidates, which is exactly what the MLLM re-decides. Showing the MLLM a four-frame storyboard of the track when it is uncertain did not help either (67.50 vs. 67.66 AP, Table 4). For LV-VIS, whose objects are mostly recognizable from one good view, the useful role of the temporal dimension is to provide that good view and to stabilize the prior.

Table 4 Design ablations and MLLM backbones (oracle setting, fixed fusion $T = 0.07$, $\alpha = 0.95$). Each variant changes one decision of the indicated reference; Δtop-1 and p are paired against that reference on held-out tracks.

Variant	val AP	held-out AP	top-1	Δtop-1	p
CtxVIS (reference)	68.51	67.88	49.92	–	–
describe, then choose (two-step)	68.03	67.66	49.37	−0.55	0.29
two-step (reference)	68.03	67.66	49.37	–	–
+ adaptive storyboard	67.85	67.50	49.37	+0.00	1.00
mask overlay instead of crop	66.38	65.67	48.15	−1.22	0.048
$K = 25$ candidates	68.20	67.87	49.28	−0.09	0.79
+ LLM option definitions	67.31	67.00	48.06	−1.31	0.002
Qwen3-VL-8B	66.49	66.33	48.95	–	–
InternVL3-8B	64.08	63.98	45.89	–	–
Qwen2.5-VL-3B	64.96	64.57	45.65	–	–

5.6 Ablations and MLLM backbones

Table 4 summarizes further design decisions, most of which were made on the describe-then-choose variant and are therefore paired against it. A green mask overlay as the visual prompt instead of the close-up crop costs 1.99 AP and significantly lowers top-1 accuracy. Neither a larger shortlist ($K = 25$) nor an adaptive four-frame storyboard changes the result, and option definitions generated by the LLM, which looked slightly better on the development split, are significantly worse on held-out videos (Sect. 6). Writing a description before answering does not help either, so CtxVIS answers directly. Among the MLLM backbones, all used with the fusion parameters selected for Qwen2.5-VL-7B, Qwen3-VL-8B comes closest (66.33 AP), while InternVL3-8B (63.98) and Qwen2.5-VL-3B (64.57) are weaker. All of them clearly improve on the prior alone (61.98 AP), so the method is not tied to one MLLM. Re-selecting the fusion parameters for each backbone on the 100 development videos changes their held-out AP by −0.98 to +1.64 points, i.e. the parameters found for one MLLM transfer reasonably to the others.

5.7 Fully automatic open-vocabulary VIS

Table 5 and Fig. 6 report the fully automatic setting. The released OV-VIS models reach 24.70 (OVFormer Swin-B) and 28.28 AP (OV-DVIS++ offline ConvNeXt-L). OV-DVIS++, trained on COCO, has a much lower AP_b than AP_n, which reflects the fine-grained base categories of LV-VIS. With SAM 2 tracklets, recognition quality translates directly into AP: CLIP gives 24.00 AP, Alpha-CLIP with eight views 31.29, and CtxVIS 36.37 AP. Recognition alone thus already exceeds the trained models by 8.09 AP. The gap to the oracle setting (68.51 AP) comes from the tracklets: SAM 2 masks fragment objects, follow parts, and include background, so that even perfect recognition could not recover the lost AP.

Table 5 Fully automatic OV-VIS on LV-VIS val (836 videos) and on the held-out videos. All numbers are measured by us with the same evaluator; trained baselines use their released weights. The SAM 3 row with ground-truth vocabulary uses the categories annotated in each video as prompts and is an upper bound for concept-prompted segmentation, not a comparable result.

		val				held-out	
Method	Training data	AP	AP_{50}	AP_b	AP_n	AP	AP_n
OVFormer R50	LV-VIS base	22.11	28.23	22.17	22.07	22.49	22.84
OVFormer Swin-B	LV-VIS base	24.70	30.76	27.14	22.92	25.61	23.99
OV-DVIS++ online R50	COCO	13.40	17.22	7.81	17.48	13.50	17.57
OV-DVIS++ online ConvNeXt-L	COCO	28.02	35.71	12.46	39.37	28.38	40.14
OV-DVIS++ offline ConvNeXt-L	COCO	28.28	36.22	12.81	39.57	28.57	40.23
SAM 2 + CLIP	none	24.00	28.94	16.87	29.20	23.99	29.24
SAM 2 + Alpha-CLIP, 3 views	none	30.92	38.14	24.07	35.92	30.85	36.33
SAM 2 + Alpha-CLIP, 8 views	none	31.29	38.54	22.74	37.52	31.33	38.13
SAM 2 + CtxVIS	none	36.37	44.33	27.55	42.81	36.42	43.27
+ SAM 3 re-segmentation	none	40.54	49.44	27.64	49.94	40.17	50.05
+ re-classify SAM 3 tracklets	none	40.86	49.24	27.28	50.76	40.81	50.91
+ re-scoring (full CtxVIS)	none	**41.89**	**50.93**	**28.46**	**51.69**	**41.31**	**51.32**
SAM 3, GT vocabulary (upper bound)	*none*	*55.52*	*69.40*	*51.66*	*58.33*	*55.63*	*58.27*

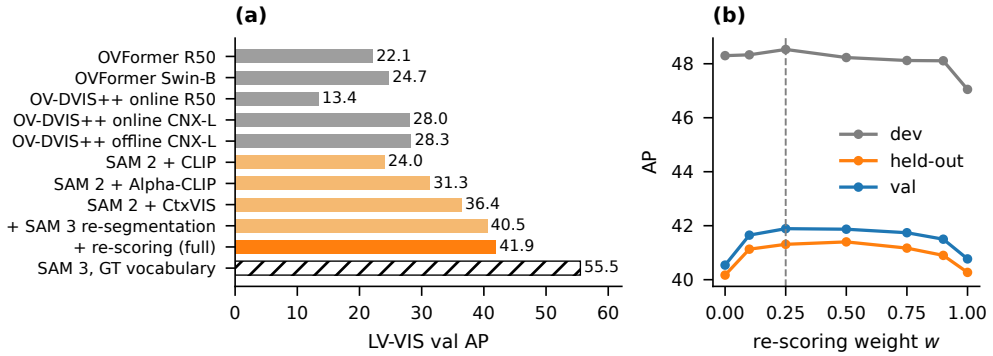


Fig. 6 Fully automatic pipeline. (a) Val AP of trained OV-VIS models and training-free pipelines; the hatched bar is the SAM 3 upper bound with the ground-truth vocabulary. (b) AP of the re-scored SAM 3 detections as a function of the re-scoring weight w (Eq. 7) on the development and held-out splits; $w = 0$ uses the SAM 3 scores only.

Vocabulary-driven re-segmentation addresses this. Prompting SAM 3 with the five best categories of each video covers 48.6% of the annotated video–category pairs and raises AP to 40.54, with the largest gain on novel categories (AP_n 42.81 → 49.94). Re-scoring the detections with the CtxVIS category evidence (Eq. 7) adds a further 1.35 AP, to 41.89 AP on val and 41.31 on held-out videos. The choice of w is not critical: any w between 0.1 and 0.9 gives 40.90–41.40 held-out AP (Fig. 6b). Re-classifying the SAM 3 tracklets with CtxVIS instead of re-scoring them is slightly worse (40.86 AP), because the tracklets were already prompted with CtxVIS categories and a second classification adds little new information. For reference, SAM 3 prompted with the ground-truth categories of each video reaches 55.52 AP. The five-category shortlist contains only 48.6% of the annotated video–category pairs, so better video-level recognition is the most direct way to close this gap. Among published methods

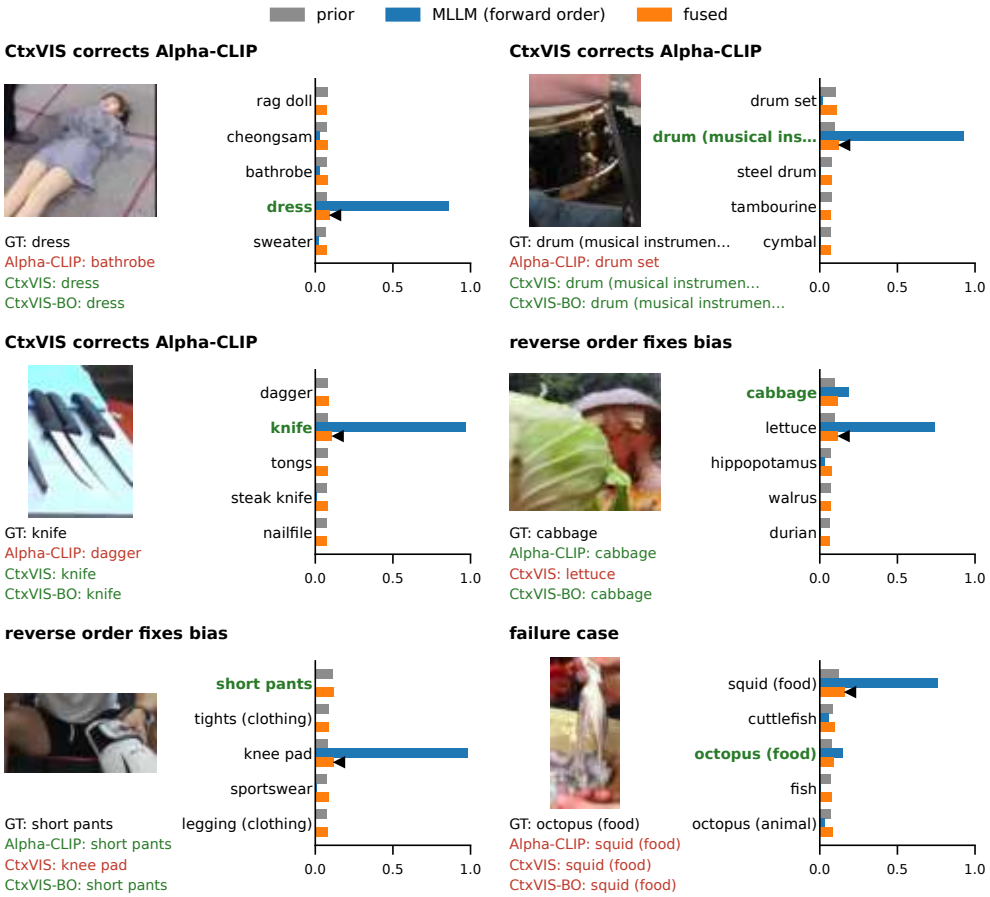


Fig. 7 Recognition examples on held-out ground-truth tracklets. For each track: the close-up crop shown to the MLLM, the ground truth, and the top-1 predictions of Alpha-CLIP, CtxVIS and CtxVIS-BO (green: correct, red: wrong). The bars show the top-5 shortlist entries with the prior (grey), the forward-order MLLM distribution (blue) and the fused distribution (orange); the triangle marks the fused decision, which can be decided by small differences because the fused distribution is flat (Sect. 5.4). The bottom-right example is a failure case.

that we did not re-run, CLIP-VIS reports 32.2 AP on LV-VIS val with a ConvNeXt-B backbone [14]; the training-free CtxVIS pipeline is 9.7 AP higher.

5.8 Qualitative results

Figure 7 shows typical cases. Alpha-CLIP confuses visually similar categories whose names the MLLM can distinguish from the crop, such as related animals or tools. The fused distribution corrects them when the prior ranks the right category among its top candidates, while the prior prevents the MLLM from following a confident but wrong answer outside the plausible set. The CtxVIS-BO examples show cases where the forward query rejects the prior's first candidate, presented as option A,

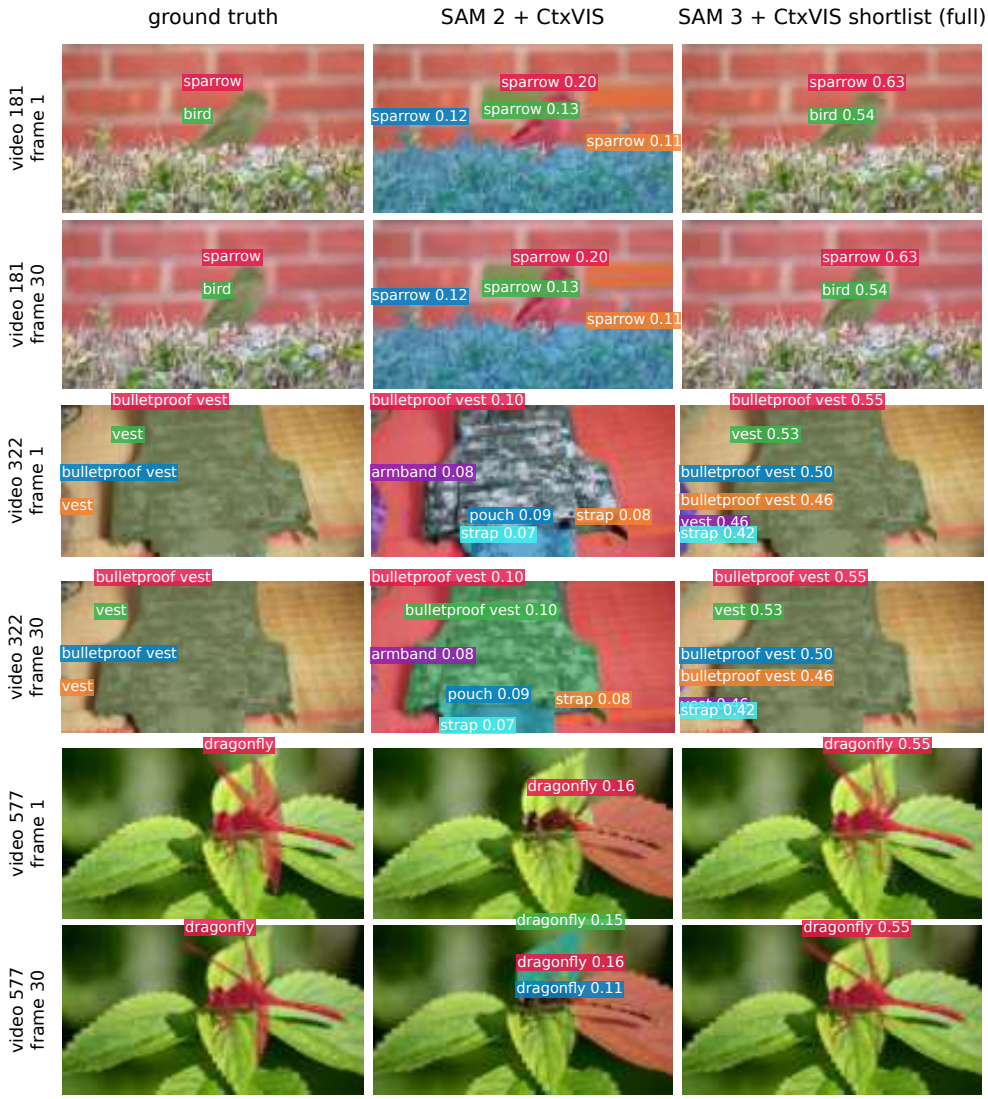


Fig. 8 Fully automatic results on held-out videos. Rows per video: ground truth, SAM 2 tracklets recognized by CtxVIS, and the final re-segmentation with SAM 3 prompted by the CtxVIS shortlist. Each method shows its $n+2$ highest-scoring detections, where n is the number of annotated objects in the video; labels give the category and score.

and the reversed query recovers it. Figure 8 illustrates why re-segmentation matters: SAM 2 tracklets split objects into parts and add background tracklets, whereas SAM 3 prompted with the correct concept returns complete instances.

Table 6 Run time on the timing split (50 held-out videos, 201 ground-truth tracks), one RTX 4090D, measured with the GPU otherwise idle. Seconds per video for each stage (tracklets: SAM 2; render: views and crops; prior: Alpha-CLIP; MLLM: multiple-choice queries; SAM 3: re-segmentation with five prompts) and peak GPU memory in GB.

Method	tracklets	render	prior	MLLM	SAM 3	total	GPU mem.
Alpha-CLIP, 3 views	–	1.6	0.9	–	–	2.4	2.5
Alpha-CLIP, 8 views (prior)	–	1.6	4.6	0.1	–	6.2	1.6
CtxVIS	–	1.6	4.6	2.3	–	8.5	17.5
CtxVIS, describe then choose	–	1.6	4.6	11.0	–	17.1	17.5
CtxVIS-BO[a]	–	1.6	4.6	4.7	–	10.8	17.5
SAM 2 + Alpha-CLIP, 8 views	23.0	4.3	3.1	0.1	–	30.5	5.0
SAM 2 + CtxVIS	23.0	4.3	13.4	9.4	–	50.2	17.4
full CtxVIS pipeline[b]	23.0	4.3	13.4	9.4	27.6	77.8	17.4

[a] One extra MLLM pass per tracklet, estimated from the measured single pass.

[b] SAM 3 measured during the validation run with two workers sharing the GPU (seconds per video and worker).

5.9 Computational cost

CtxVIS trades computation for accuracy without training (Table 6). With ground-truth tracklets, the region prior costs 4.6 s per video and the MLLM query about 0.58 s per tracklet, i.e. 8.5 s per video; CtxVIS-BO doubles the MLLM part. In the fully automatic pipeline, SAM 2 tracklet generation adds 23.0 s per video, and the SAM 3 re-segmentation with five prompts took 27.6 s per video and worker in our validation run (two workers sharing one GPU, 5.5 s per prompt). The pipeline is therefore an offline method, considerably slower than trained single-model VIS networks; its components are independent per tracklet and per prompt and parallelize trivially.

6 Limitations

One benchmark.

LV-VIS is the only large-vocabulary VIS benchmark with exhaustive labels and public validation annotations. We reduced the risk of tuning to it with a development split and paired tests, but cannot rule out that some conclusions, e.g. that additional views add little inside the pipeline, are specific to its mostly short, object-centric videos.

Training data of the foundation models.

SAM 3 is trained on a very large concept-segmentation corpus, and Alpha-CLIP and the MLLMs on web-scale data. Although none of the models is trained on LV-VIS by us, we cannot exclude an overlap of their training images or concepts with LV-VIS, and "novel" categories are novel only with respect to LV-VIS training, not to the foundation models. The comparison with trained OV-VIS models should be read with this in mind; the comparison within the oracle setting uses the same foundation models for all methods.

Selection and multiple comparisons.

The fusion parameters and the re-scoring weight were selected on the development split only. Several analysis-driven variants (option orders and their combinations, the second view, alternative priors) were evaluated on the same held-out videos; we report all of them, use the single-query CtxVIS as the main method, and present CtxVIS-BO as a modest improvement. One variant illustrates the risk: LLM-generated option definitions looked slightly better on the development split (15 vs. 10 tracks, $p = 0.42$) but were significantly worse on held-out videos (70 vs. 113, $p = 0.002$).

Cost and tracklets.

An MLLM forward pass per tracklet and a SAM 3 pass per prompt make the pipeline slow compared with trained VIS models. The re-segmentation stage also depends on SAM 2 tracklets to form the shortlist and inherits their misses.

7 Conclusion

We presented CtxVIS, a training-free framework for open-vocabulary video instance segmentation that uses a multi-view region prior to shortlist categories, an MLLM to rerank them with one multiple-choice query, and a tempered product of experts to combine the two. A careful analysis showed that reranking, not candidate recall, limits recognition; that the MLLM's over-confidence calls for a dominant, retempered prior; that better temporal pooling of the prior is largely absorbed by the reranker; and that the MLLM's preference for late options is cancelled by querying in two orders. The recognized categories of a video drive concept-prompted re-segmentation with SAM 3, giving 41.89 AP on LV-VIS val, well above trained OV-VIS models evaluated under the same protocol. Future work includes reducing the cost by distilling the reranker, extending the evaluation to other large-vocabulary video benchmarks, and closing the loop by re-segmenting with categories that the re-segmented masks suggest.

Statements and Declarations

- **Funding:** No funding was received for conducting this study.
- **Competing interests:** The authors have no competing interests to declare that are relevant to the content of this article.
- **Ethics approval and consent to participate:** Not applicable. The study uses only the public LV-VIS benchmark.
- **Consent for publication:** Not applicable.
- **Data availability:** LV-VIS is publicly available from its authors [7]. The development, held-out and timing split lists, all per-track predictions and the evaluation outputs used in this paper are available from the corresponding author on reasonable request and will be released with the code.
- **Materials availability:** Not applicable.
- **Code availability:** The code, configuration files and scripts that reproduce every table and figure will be released upon publication.

- **Use of AI tools:** An AI assistant supported the writing of evaluation and plotting scripts and the drafting and language editing of the text (see Sect. 4). The authors checked all results and conclusions and take full responsibility for the article.
- **Author contribution:** Gen Li and Mingyong Zhang contributed equally to this work. Conceptualization: Gen Li, Wenfeng Zhang; Methodology: Gen Li, Mingyong Zhang; Software: Gen Li, Mingyong Zhang; Investigation: Gen Li, Mingyong Zhang; Formal analysis: Gen Li; Validation: Mingyong Zhang, Gen Li; Data curation: Mingyong Zhang; Visualization: Gen Li; Writing – original draft: Gen Li; Writing – review and editing: Mingyong Zhang, Wenfeng Zhang; Resources: Wenfeng Zhang; Supervision: Wenfeng Zhang; Project administration: Wenfeng Zhang. All authors read and approved the final manuscript.

References

[1] Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5188–5197 (2019)

[2] Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H.S., Bai, S.: Occluded video instance segmentation: A benchmark. International Journal of Computer Vision **130**, 2022–2039 (2022)

[3] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1299 (2022)

[4] Huang, D.-A., Yu, Z., Anandkumar, A.: MinVIS: A minimal video instance segmentation framework without video-based training. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

[5] Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)

[6] Zhang, T., Tian, X., Wu, Y., Ji, S., Wang, X., Zhang, Y., Wan, P.: DVIS: Decoupled video instance segmentation framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1282–1291 (2023)

[7] Wang, H., Yan, C., Wang, S., Jiang, X., Tang, X., Hu, Y., Xie, W., Gavves, E.: Towards open-vocabulary video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4057–4066 (2023)

[8] Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: A large-scale benchmark for tracking any object. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 436–454 (2020)

[9] Athar, A., Luiten, J., Voigtlaender, P., Khurana, T., Dave, A., Leibe, B., Ramanan, D.: BURST: A benchmark for unifying object recognition, segmentation and tracking in video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1674–1683 (2023)

[10] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)

[11] Guo, P., Huang, T., He, P., Liu, X., Xiao, T., Chen, Z., Zhang, W.: OpenVIS: Open-vocabulary video instance segmentation. arXiv preprint arXiv:2305.16835 (2023)

[12] Cheng, Z., Li, K., Li, H., Jin, P., Liu, C., Zheng, X., Ji, R., Chen, J.: Instance Brownian bridge as texts for open-vocabulary video instance segmentation. arXiv preprint arXiv:2401.09732 (2024)

[13] Fang, H., Wu, P., Li, Y., Zhang, X., Lu, X.: Unified embedding alignment for open-vocabulary video instance segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 225–241 (2024)

[14] Zhu, W., Cao, J., Xie, J., Yang, S., Pang, Y.: CLIP-VIS: Adapting CLIP for open-vocabulary video instance segmentation. IEEE Transactions on Circuits and Systems for Video Technology **35**(2), 1098–1110 (2025) https://doi.org/10.1109/TCSVT.2024.3474698

[15] Zhang, T., Tian, X., Zhou, Y., et al.: DVIS++: Improved decoupled framework for universal video segmentation. arXiv preprint arXiv:2312.13305 (2023)

[16] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: International Conference on Learning Representations (ICLR) (2025)

[17] Carion, N., et al.: SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)

[18] Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-CLIP: A CLIP model focusing on wherever you want. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13019–13029 (2024)

[19] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)

[20] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)

[21] Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)

[22] Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: OpenMask3D: Open-vocabulary 3D instance segmentation. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)

[23] Zhou, M., He, C., Wang, R., Chen, X.: OV3D-CG: Open-vocabulary 3D instance segmentation with contextual guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5305–5314 (2025)

[24] Zhang, Y., Unell, A., Wang, X., Ghosh, D., Su, Y., Schmidt, L., Yeung-Levy, S.: Why are visually-grounded language models bad at image classification? In: Advances in Neural Information Processing Systems (NeurIPS) (2024)

[25] Wang, J., Cao, H., Sun, C., Huang, Z., Zhang, Y.: Motion perception-driven multimodal self-supervised video object segmentation. The Visual Computer **41** (2025) https://doi.org/10.1007/s00371-024-03597-8

[26] Qu, D., Zhao, H., Zhou, M.: Unsupervised video object segmentation with mask transformer: Boosting accuracy and efficiency through feature fusion. The Visual Computer **41**(10), 7509–7520 (2025) https://doi.org/10.1007/s00371-025-03820-0

[27] Yi, Y., Gui, Y., Liu, Z.: Boosting memory network for video object segmentation in complex scenes. The Visual Computer **41**(9), 6571–6585 (2025)

[28] Gu, X., Lin, T.-Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (ICLR) (2022)

[29] Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 350–368 (2022)

[30] Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 728–755 (2022)

[31] Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L.,

Zhang, L., Hwang, J.-N., Chang, K.-W., Gao, J.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10965–10975 (2022)

[32] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 38–55 (2024)

[33] Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted CLIP. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7061–7070 (2023)

[34] Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2955–2966 (2023)

[35] Yu, Q., He, J., Deng, X., Shen, X., Chen, L.-C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)

[36] Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)

[37] Muttal, C., Giri, C., Mulla, M.N., Dhane, R.: Semantic bridge-guided training-free open-vocabulary segmentation for fashion visual understanding. The Visual Computer **42**, 520 (2026) https://doi.org/10.1007/s00371-026-04732-3

[38] Chen, T., Liang, S., Xiong, Y., *et al.*: Enhancing open-vocabulary scene understanding via push–pull alignment in Gaussian splatting. The Visual Computer **42**, 38 (2026) https://doi.org/10.1007/s00371-025-04219-7

[39] Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does CLIP know about a red circle? visual prompt engineering for VLMs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11987–11997 (2023)

[40] Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. arXiv preprint arXiv:2310.11441 (2023)

[41] You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. In: International Conference on Learning Representations (ICLR) (2024)

[42] Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 28202–28211 (2024)

[43] Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15691–15701 (2023)

[44] Zheng, C., Zhou, H., Meng, F., Zhou, J., Huang, M.: Large language models are not robust multiple choice selectors. In: International Conference on Learning Representations (ICLR) (2024)

[45] Pezeshkpour, P., Hruschka, E.: Large language models sensitivity to the order of options in multiple-choice questions. In: Findings of the Association for Computational Linguistics: NAACL 2024, pp. 2006–2017 (2024)

[46] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., Sui, Z.: Large language models are not fair evaluators. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 9440–9450 (2024)

[47] Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 12697–12706 (2021)

[48] Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024)

[49] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(3), 226–239 (1998)

[50] Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation **14**(8), 1771–1800 (2002)

[51] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 1321–1330 (2017)

[52] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026 (2023)

[53] Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.-Y.: Tracking anything with

decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1316–1326 (2023)

[54] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)

[55] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)

[56] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)

Author biographies

Gen Li received a master's degree and is an Engineer (intermediate professional title) at the Guangdong Provincial Second Hospital of Traditional Chinese Medicine (Guangdong Provincial Engineering Technology Research Institute of Traditional Chinese Medicine), Guangzhou, China, and is affiliated with the Guangdong Provincial Key Laboratory of Research and Development in Traditional Chinese Medicine. The research interests of Gen Li are in computer vision, in particular open-vocabulary video understanding, video instance segmentation and visual recognition with multimodal foundation models.

Mingyong Zhang holds a bachelor's degree and is an Engineer (intermediate professional title) at the Guangdong Provincial Second Hospital of Traditional Chinese Medicine (Guangdong Provincial Engineering Technology Research Institute of Traditional Chinese Medicine) and the Guangdong Provincial Key Laboratory of Research and Development in Traditional Chinese Medicine, Guangzhou, China. The research field of Mingyong Zhang is computer science and technology.

Wenfeng Zhang holds a bachelor's degree and is a Senior Engineer at the Guangdong Provincial Second Hospital of Traditional Chinese Medicine (Guangdong Provincial Engineering Technology Research Institute of Traditional Chinese Medicine) and the Guangdong Provincial Key Laboratory of Research and Development in Traditional Chinese Medicine, Guangzhou, China. The research field of Wenfeng Zhang is computer science and technology, and Wenfeng Zhang is the corresponding author of this article.