MiNQVIS: Mitigating Noisy Queries for Robust Online Video Instance Segmentation

Jie Qiao[1], Jianxu Chen[2], and Xiaowei Xu[2]✉

[1] School of Computer Science, Wuhan University, Wuhan, China
jie.qiao@whu.edu.cn
[2] Leibniz-Institut für Analytische Wissenschaften, Germany
jianxu.chen@isas.de, xiao.wei.xu@foxmail.com

Abstract. Online Video Instance Segmentation (VIS) relies on per-frame instance queries for causal detection, segmentation, and temporal association. However, in realistic videos with occlusion, blur, and clutter, these predictions often become noisy—exhibiting false positives, missed instances, or inaccurate masks. Such noisy queries not only destabilize cross-frame association but also contaminate contrastive supervision and propagate drift through memory-based matching. Despite advances in recent online VIS methods, none explicitly identify or address noisy query contamination, leaving a fundamental weakness unexamined. We introduce MiNQVIS, the first noise-aware framework for online VIS that stabilizes both training and inference. Our approach consists of three complementary components: (1) query-level gating that filters unreliable predictions using classification-guided thresholding and embedding-norm stabilization; (2) video-level prototypical contrastive learning that aggregates only reliable matches into stable, noise-robust prototypes; (3) dual-memory association that fuses momentum embeddings with history-best retrieval to prevent drift from corrupted frames. Experiments on YouTubeVIS 2019/2021 and OVIS show consistent improvements over strong online VIS baselines, with large gains in challenging scenes. Project Page: https://nothing898.github.io/MiNQVIS.

Keywords: Video instance segmentation · Contrastive learning · Temporal Association · Representation Learning

1 Introduction

Video Instance Segmentation (VIS) aims to detect, segment, and temporally associate object instances in videos, and forms a critical component of modern video-centric perception systems such as autonomous driving, augmented and mixed reality, and embodied intelligence [18]. In this context, online VIS [8, 10, 11, 17–20], which processes streaming inputs in a strictly causal manner, has gained substantial attention due to its computational efficiency and suitability for real-world deployment.

✉ Corresponding author.

Recent progress [8, 10, 11, 17, 20] has been driven primarily by query-based online VIS frameworks, which unify detection, segmentation, and temporal association through learnable instance queries. While achieving strong empirical performance, these methods fundamentally rely on the assumption that per-frame predicted queries provide reliable cues for cross-frame association and contrastive objectives used during training [16, 17, 22]. In realistic video scenarios involving severe occlusion, motion blur, background clutter, or rapid appearance changes, this assumption is frequently violated [15, 17, 18].

The resulting noisy queries—including false positives, missed detections, and poorly localized predictions—often produce corrupted embeddings that distort similarity-based matching and learning signals across frames. This leads to three fundamental challenges: (1) **unstable cross-frame association**, as noisy embeddings distort similarity-based matching; (2) **contaminated contrastive supervision**, since corrupted embeddings participate in positive/negative pair construction; and (3) **temporal drift in memory-based tracking**, where erroneous updates propagate through momentum-based embeddings. Although recent methods such as CAVIS [11], VISAGE [10], and CTVIS [20] improve temporal consistency, none explicitly address the root cause—noisy query contamination—which remains a primary bottleneck in online VIS robustness [15, 17].

This observation motivates revisiting online VIS from a noise-aware perspective. Rather than solely improving architectural capacity, we argue that robust online tracking requires stabilizing the **temporal information flow** and regulating the **query-driven influence on association and supervision**. Our investigation leads to three central intuitions that directly drive the design of our framework. **First**, predicted queries vary substantially in reliability; low-confidence or irregular-norm embeddings often correspond to background confusion and should be suppressed before influencing downstream modules. This motivates the introduction of a **query-level gating mechanism**. **Second**, single-frame embeddings are particularly unreliable under challenging conditions, suggesting that contrastive learning should leverage temporally aggregated and stable signals rather than per-frame features. This motivates our **video-level prototype contrastive supervision**. **Third**, memory representations used for online association should be resilient to short-term feature corruption; relying solely on momentum updates makes the system vulnerable to drift. This insight motivates a **dual-memory association strategy** that jointly exploits momentum-based embeddings and history-best retrieval.

Guided by these principles, we introduce MiNQVIS, a unified noisy-query mitigation framework for online VIS comprising three complementary components: (1) **query-level gating** through classification-guided thresholding and embedding-norm stabilization; (2) **video-level prototypical contrastive learning** (**VPCL**) constructed from reliable cross-frame matches; and (3) **dual-memory fusion** combining momentum embeddings with history-best retrieval for drift-resistant association. Extensive experiments on YouTubeVIS 2019/2021 [18] and OVIS [15] demonstrate consistent improvements over strong online baselines, with significant gains under heavy occlusion and motion degradation.

Our contributions are as follows:

- We systematically analyze noisy instance queries and identify them as a key source of instability in online VIS, which can corrupt temporal association and contrastive supervision across frames.
- We propose a noise-aware framework for online VIS consisting of three components: query-level gating for filtering unreliable queries, VPCL for constructing stable instance representations, and dual-memory fusion for drift-resistant online association.
- Extensive experiments on YouTubeVIS 2019/2021 and OVIS demonstrate consistent improvements over strong online VIS baselines under both ResNet-50 and Swin-L backbones.

2 Related Work

2.1 Online Video Instance Segmentation

Online VIS [8, 10, 11, 17–20] aims to detect, segment, and associate object instances in a causal manner without accessing future frames. Early approaches [18, 19] extend image-based instance segmentation with heuristic tracking modules, while recent methods [8, 10, 11, 17, 20] rely on learnable query-based representations [2, 4, 24] for stronger temporal consistency. IDOL [17] introduces contrastive learning to improve cross-frame association via instance-level embeddings. VISAGE [10] enhances appearance modeling by refining visual features and query representations. CTVIS [20] encourages training–inference consistency through momentum embeddings and cross-frame interactions. CAVIS [11] incorporates context-aware matching and prototypical contrastive objectives to stabilize association. Although effective, all these methods share a critical assumption: per-frame predicted queries are sufficiently reliable to serve as association cues and training signals. In challenging videos with occlusion, motion blur, or background clutter, this assumption breaks down (**unstable cross-frame association**), which is a direct consequence of **noisy queries** frequently produced under challenging conditions. None of the above explicitly identify or handle the issue of noisy queries, which can propagate errors through association and training, forming a fundamental limitation of existing online VIS systems.

2.2 Contrastive Learning for Temporal Association

Contrastive learning has become a prevalent mechanism for improving temporal association in VIS. IDOL [17] leverages cross-video contrastive supervision to enhance discriminability, while CAVIS [11] and VISAGE [10] adopt prototypical or localized contrastive objectives to strengthen cross-frame correspondence. More broadly, instance-level InfoNCE-based supervision has been widely applied in video recognition, tracking, and re-identification tasks, where embedding similarity plays a central role in aligning temporal representations. However, these methods uniformly assume clean and reliable embeddings. When

noisy queries participate in constructing positive or negative pairs, contrastive objectives are easily corrupted, leading to unstable or contradictory supervision (**contaminated contrastive supervision**). To our knowledge, prior VIS work does not explicitly address robust contrastive learning under noisy embeddings, nor does it provide mechanisms to ensure that only trustworthy temporal matches contribute to prototype formation.

2.3 Memory-based Tracking and Drift

Memory-based association is widely used in online VIS and related video tasks. IDOL [17] introduces a memory bank for online instance association, and similar memory-based designs are adopted in later methods such as CTVIS [20] and VISAGE [10] to improve temporal association. In particular, momentum memory has been used in CTVIS and related contrastive tracking frameworks to maintain an evolving representation of each instance over time. However, momentum memory suffers from a well-known limitation: drift accumulation (**temporal drift in memory-based tracking**). When several consecutive frames produce corrupted embeddings, the memory state may be gradually overwritten, causing errors to persist across long temporal spans. Existing online VIS approaches rely solely on such momentum-updated representations and lack mechanisms to counteract this drift. To address this issue, we introduce a dual-memory design that combines momentum embeddings for short-term adaptability with a history-best retrieval from a cache of historical embeddings. By jointly leveraging these two complementary memories, our approach improves robustness to short-term corruption and stabilizes long-term association.

3 Method

3.1 Overview

As illustrated in Fig. 1, we follow the standard online VIS pipeline built on Mask2Former [4]. Given a video $\{I_t\}_{t=1}^{T}$, the model produces N_q instance queries per frame with classification logits, masks, and embeddings. During training, Hungarian matching assigns queries to ground-truth instances to provide supervision.

On top of this baseline, we introduce three complementary components to improve robustness: (1) **Query-level gating**, which filters unreliable queries using classification confidence and ℓ_2-normalized embeddings; (2) **VPCL**, which aggregates reliable cross-frame embeddings into video-level prototypes for contrastive supervision during training; (3) **Dual-memory fusion association**, which combines momentum embeddings and history-best retrieval to reduce temporal drift during inference.

Query-level gating is applied in both training and inference. During training, gated embeddings are used to construct contrastive pairs and optimize the VPCL objective. During inference, they are passed to the dual-memory module to produce stable track identities.

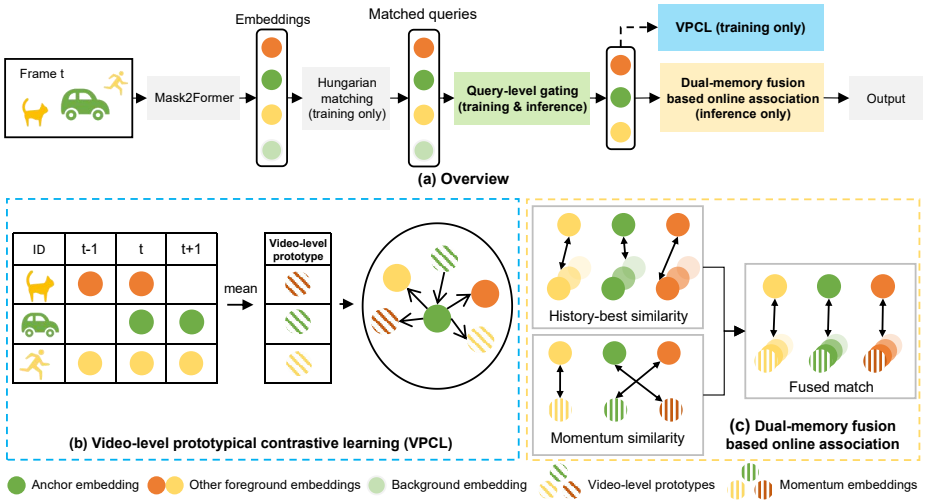

Fig. 1: Overview of MiNQVIS. (a) Mask2Former produces per-frame queries and embeddings. During training, Hungarian matching assigns queries to ground-truth instances, and the matched queries are filtered by query-level gating (applied in both training and inference). The gated embeddings are used for VPCL training and online association. (b) VPCL aggregates reliable cross-frame embeddings to form video-level prototypes for contrastive supervision. (c) During inference, the dual-memory module fuses momentum embeddings and history-best retrieval to perform robust online association and maintain consistent identities.

While gating mainly suppresses low-confidence noisy queries such as false positives and fragmented masks, VPCL and dual-memory further mitigate such residual noise by stabilizing representation learning and temporal association during training and inference.

The overall training objective combines the standard Mask2Former segmentation losses $\mathcal{L}_{\text{seg}}$, including the classification loss $\mathcal{L}_{\text{cls}}$, cross-entropy mask loss $\mathcal{L}_{\text{ce}}$, and dice loss $\mathcal{L}_{\text{dice}}$, with the proposed gating regularization loss $\mathcal{L}_{\text{gate}}$ (Sec. 3.2) and the video-level prototypical contrastive loss $\mathcal{L}_{\text{vpcl}}$ (Sec. 3.3). Formally, the training objective is

$$\begin{aligned} \mathcal{L}_{\text{seg}} &= \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}, \\ \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{seg}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_{\text{vpcl}}\mathcal{L}_{\text{vpcl}}. \end{aligned} \tag{1}$$

3.2 Query-level gating

Online VIS models produce a large number of per-frame instance queries, many of which correspond to background regions or ambiguous detections under occlusion, blur, or clutter. Such unreliable queries can introduce noisy gradients

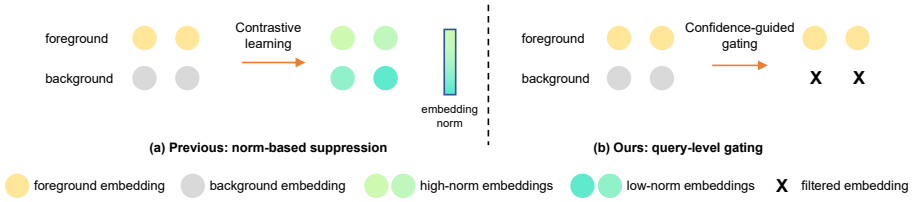

Fig. 2: Illustration of query-level gating. (a) With unnormalized inner-product similarity, background queries may be suppressed implicitly via norm shrinkage, leading to unstable matching scores. (b) Our method explicitly suppresses unreliable queries using a classification-guided gate and stabilizes similarity via ℓ_2 normalization.

during contrastive training and lead to spurious matches during online association. To address this issue, we introduce a lightweight *query-level gating* module that filters unreliable queries based on classification confidence and stabilizes similarity computation via ℓ_2-normalized embeddings. Unlike simple confidence thresholding used in standard detectors, our gating mechanism is integrated with contrastive learning and temporal association, ensuring that unreliable queries are consistently filtered during both training and inference.

Confidence-guided gating. Let $\mathbf{z}_{t,j} \in \mathbb{R}^{C+1}$ denote the classification logits of query j at frame t, where the $(C+1)$-th entry corresponds to the background class. We compute foreground and background probabilities as

$$p^j_{t,\text{bg}} = \text{softmax}(\mathbf{z}_{t,j})_{C+1}, \qquad p^j_{t,\text{fg}} = 1 - p^j_{t,\text{bg}}. \tag{2}$$

A query is retained only if $p^j_{t,\text{fg}} > \gamma$, where γ is a fixed confidence threshold determined empirically. This gate removes low-confidence queries from both contrastive supervision and online matching.

To further separate foreground and background predictions, we introduce a gating regularization loss $\mathcal{L}_{\text{gate}}$. Let $\mathcal{S}^+$ and $\mathcal{S}^-$ denote the sets of matched (foreground) and unmatched queries obtained from Hungarian assignment. Given γ and a margin μ, we define

$$p^\star_{\text{fg}} = \gamma + \mu, \qquad p^\star_{\text{bg}} = 1 - \gamma + \mu,$$

where μ is a small positive margin that enforces a confidence gap between foreground and background predictions. In all experiments, we set $\mu = 0.1\gamma$. The gating threshold γ is not highly sensitive and we analyze its influence in Sec. 4.4. For matched queries we encourage $p_{t,\text{fg}}$ to exceed $p^\star_{\text{fg}}$, while for unmatched queries we encourage $p_{t,\text{bg}}$ to exceed $p^\star_{\text{bg}}$. This is implemented using a log-domain hinge loss:

$$\mathcal{L}_{\text{gate}} = \frac{1}{|\mathcal{S}^+|} \sum_{(t,j)\in\mathcal{S}^+} \Big[\log p^\star_{\text{fg}} - \log(p^j_{t,\text{fg}} + \epsilon) \Big]_+ + \frac{1}{|\mathcal{S}^-|} \sum_{(t,j)\in\mathcal{S}^-} \Big[\log p^\star_{\text{bg}} - \log(p^j_{t,\text{bg}} + \epsilon) \Big]_+, \tag{3}$$

where $[x]_+ = \max(x, 0)$ and ϵ is a small constant for numerical stability.

Embedding normalization. Similarity scores computed using raw inner products are sensitive to feature magnitude and may couple similarity with embedding norms. To decouple similarity from feature magnitude, we normalize the query embedding e_t^j as

$$\hat{e}_t^j = \frac{e_t^j}{\|e_t^j\|_2 + \epsilon}, \tag{4}$$

and use cosine similarity for both contrastive learning and temporal association.

3.3 Video-level prototypical contrastive learning (VPCL)

Single-frame embeddings are often unreliable under occlusion and motion blur. To obtain more stable supervision, we construct *video-level prototypes* by aggregating reliable cross-frame matches within a training clip and use them for contrastive learning. Compared with frame-level contrastive learning, prototype aggregation reduces the influence of noisy frames and provides more stable identity supervision across the clip.

Video-level prototypes. For each instance k, we compute a prototype by averaging the normalized embeddings matched to that instance across the clip:

$$\hat{p}_k = \frac{p_k}{\|p_k\|_2}, \qquad p_k = \frac{1}{|\mathcal{E}_k|} \sum_{\hat{e} \in \mathcal{E}_k} \hat{e}, \qquad \mathcal{E}_k = \{\hat{e}_{t,j(t,k)} \mid (t,k) \text{ is matched}\}, \tag{5}$$

where $|\mathcal{E}_k|$ denotes the number of embeddings matched to instance k. These prototypes serve as stable temporal references and reduce the influence of noisy or ambiguous frames, resulting in more consistent identity representations.

Contrastive pair construction. For each matched pair (t,k) from Hungarian assignment, the normalized embedding $\hat{e}_{t,j(t,k)}$ serves as the anchor e, and the corresponding prototype $\hat{p}_k$ is used as the positive e^+. Negatives are drawn from two sources: (i) high-confidence foreground embeddings from the previous frame and (ii) prototypes of other instances in the same clip:

$$\mathcal{N}_{t,k} = \mathcal{N}_{t,k}^{\text{prev}} \cup \mathcal{N}_{t,k}^{\text{proto}}, \tag{6}$$

where $\mathcal{N}_{t,k}^{\text{prev}}$ denotes foreground embeddings from the previous frame, and $\mathcal{N}_{t,k}^{\text{proto}}$ denotes prototypes of other instances within the clip. We denote the set of negatives as $\{e_k^-\}$ and use the following contrastive objective:

$$\mathcal{L}_{\text{vpcl}}(e, e^+, \{e_k^-\}) = \log\left[1 + \sum_k \exp\big(\text{sim}(e, e_k^-) - \text{sim}(e, e^+)\big)\right], \tag{7}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. Since embeddings are ℓ_2-normalized, cosine similarities are naturally bounded; therefore we apply the contrastive objective directly without introducing an additional temperature parameter.

3.4 Dual-memory fusion based online association

Online association becomes unreliable when current-frame embeddings are degraded by occlusion or abrupt appearance changes. To improve robustness, we introduce a *dual-memory fusion* strategy that combines (i) a momentum embedding capturing short-term continuity and (ii) a history-best retrieval that retrieves the best-matching historical embedding from a track cache.

Given query–tracklet cosine similarities, raw scores may not be directly comparable across queries or tracklets. We therefore calibrate the matching scores using a bi-directional softmax (query-to-tracklet and tracklet-to-query normalization), which improves score comparability and matching stability.

Track memory. For each active tracklet k, we maintain a momentum embedding $m_k \in \mathbb{R}^d$ and a small cache of historical embeddings $\mathcal{R}_k = \{r_{k,1}, \ldots, r_{k,N_k}\}$. When instance k is matched at frame t, the momentum embedding is updated via exponential moving average:

$$m_k \leftarrow (1-\alpha)\, m_k + \alpha\, \hat{e}_{t,j(t,k)}, \tag{8}$$

where $\alpha \in (0,1)$, $\hat{e}_{t,j(t,k)}$ denotes the embedding of the query matched to tracklet k at frame t, and all embeddings are ℓ_2-normalized.

Dual-memory similarity. Given a query embedding $\hat{e}_{t,q}$, we first compute two *raw* similarities: (i) the *history-best similarity*, obtained by matching against the cache and selecting the maximum:

$$s_t^{\text{best}}(q,k) = \max_{r \in \mathcal{R}_k} \langle \hat{e}_{t,q}, \hat{r} \rangle\,, \tag{9}$$

where $\hat{r}$ denotes the ℓ_2-normalized historical embedding, q indexes a query in frame t and k indexes an active tracklet, and (ii) the *momentum similarity*,

$$s_t^{\text{mom}}(q,k) = \langle \hat{e}_{t,q}, \hat{m}_k \rangle\,, \tag{10}$$

where the momentum embedding is ℓ_2-normalized after each update. Instead of directly fusing these raw cosine similarities, we convert each similarity matrix into a *normalized matching score* using the bi-directional softmax.

Concretely, for $u \in \{\text{best}, \text{mom}\}$, we define

$$f_t^u(q,k) = \frac{1}{2}\left(\frac{\exp\big(s_t^u(q,k)\big)}{\sum_{k'} \exp\big(s_t^u(q,k')\big)} + \frac{\exp\big(s_t^u(q,k)\big)}{\sum_{q'} \exp\big(s_t^u(q',k)\big)} \right), \tag{11}$$

where the first term normalizes over tracklets for each query ($q \to k$) and the second term normalizes over queries for each tracklet ($k \to q$).

Table 1: Comparison on YouTubeVIS 2019/2021 and OVIS.

	Methods	YTVIS19					YTVIS21					OVIS				
		AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	MaskTrack R-CNN [18]	30.3	51.1	32.6	31.0	35.5	28.6	48.9	29.6	26.5	33.8	10.8	25.3	8.5	7.9	14.9
	SipMask [1]	33.7	54.1	35.8	35.4	40.1	31.7	52.5	34.0	30.8	37.8	10.2	24.7	7.8	7.9	15.8
	CrossVIS [19]	36.3	56.8	38.9	35.6	40.7	34.2	54.4	37.9	30.4	38.2	14.9	32.7	12.1	10.3	19.8
	IFC [9]	41.2	65.1	44.6	42.3	49.6	35.2	55.9	37.7	32.6	42.9	13.1	27.8	11.6	9.4	23.9
	Mask2Former-VIS [3]	46.4	68.0	50.0	-	-	40.6	60.9	41.8	-	-	17.3	37.3	15.1	10.5	23.5
	SeqFormer [16]	47.4	69.8	51.8	45.5	54.8	40.5	62.4	43.7	36.1	48.1	15.1	31.9	13.8	10.4	27.1
	MinVIS [8]	47.4	69.0	52.1	45.7	55.7	44.2	66.0	48.1	39.2	51.7	25.0	45.5	24.0	13.9	29.7
	VITA [7]	49.8	72.6	54.5	49.4	61.0	45.7	67.4	49.5	40.9	53.6	19.6	41.2	17.4	11.7	26.0
	IDOL [17]	49.5	74.0	52.9	47.7	58.7	43.9	68.0	49.6	38.0	50.9	28.2	51.0	28.0	14.5	38.6
	GENVIS [6]	50.0	71.5	54.6	49.5	59.7	47.1	67.5	51.5	41.6	54.7	35.8	60.8	36.2	16.3	39.6
	DVIS [23]	51.2	73.8	57.1	47.2	59.3	46.4	68.4	49.6	39.7	53.5	30.2	55.0	30.5	14.5	37.3
	TCOVIS [12]	52.3	73.5	57.6	49.8	60.2	49.5	71.2	53.8	41.3	55.9	35.3	60.7	36.6	15.7	39.5
	CTVIS [20]	55.1	78.2	59.1	**51.9**	63.2	50.1	73.7	54.7	41.8	**59.5**	35.5	60.8	34.9	16.1	41.9
	VISAGE [10]	55.1	78.1	60.6	51.0	62.3	51.6	73.8	56.1	43.6	59.3	36.2	60.3	35.3	16.1	40.3
	CAVIS [11]	**55.7**	78.3	**61.7**	51.5	**63.3**	50.5	74.1	54.9	42.6	58.5	37.6	**63.4**	38.2	16.5	43.5
	Ours	55.6	**78.5**	61.0	50.5	61.4	**51.7**	**74.4**	**56.6**	**43.9**	57.7	**38.2**	**63.4**	**38.7**	**17.6**	**45.4**
Swin-L	SeqFormer [16]	59.3	82.1	66.4	51.7	64.4	51.8	74.6	58.2	42.8	58.1	-	-	-	-	-
	Mask2Former-VIS [3]	60.4	84.4	67.0	-	-	52.6	76.4	57.2	-	-	25.8	46.5	24.4	13.7	32.2
	MinVIS [8]	61.6	83.3	68.6	54.8	66.6	55.3	76.6	62.0	45.9	60.8	39.4	61.5	41.3	18.1	43.3
	VITA [7]	63.0	86.9	67.9	56.3	68.1	57.5	80.6	61.0	47.7	62.6	27.7	51.9	24.9	14.9	33.0
	IDOL [17]	64.3	87.5	71.0	55.6	69.1	56.1	80.8	63.5	45.0	60.1	40.0	63.1	40.5	17.6	46.4
	GENVIS [6]	64.0	84.9	68.3	56.1	69.4	59.6	80.9	65.8	48.7	65.0	45.2	69.1	48.4	19.1	48.6
	DVIS [23]	63.9	87.2	70.4	56.2	69.0	58.7	80.4	66.6	47.5	64.6	45.9	71.1	48.3	18.5	51.5
	TCOVIS [12]	64.1	86.6	69.5	55.8	69.0	61.3	82.9	68.0	48.6	65.1	46.7	70.9	49.5	19.1	50.8
	CTVIS [20]	65.6	87.7	72.2	56.5	70.4	61.2	84.0	68.8	48.0	65.8	46.9	71.5	47.5	19.1	52.1
	CAVIS [11]	66.0	**89.5**	73.3	**56.8**	**71.4**	61.1	**84.1**	**69.2**	48.2	**66.3**	48.6	**74.0**	52.5	19.5	53.3
	Ours	**66.2**	87.7	**73.8**	56.1	70.0	**61.6**	**84.1**	68.4	**49.0**	65.6	**49.1**	73.9	**52.6**	**19.7**	**53.9**

Fusion and assignment. We then fuse the two normalized scores:

$$S_t(q,k) = w\, f_t^{\text{best}}(q,k) + (1-w)\, f_t^{\text{mom}}(q,k), \qquad w \in [0,1], \tag{12}$$

and perform one-to-one matching by maximizing $S_t(q,k)$ over all query–tracklet pairs. A match is accepted when $S_t(q,k) > \delta$, where δ is a score threshold. In all experiments we use fixed default values $w = 0.5$ and $\delta = 0.3$, as performance is relatively insensitive to these parameters. The fusion balances short-term continuity and reliable historical observations, improving robustness to occlusion and appearance variation.

4 Experiments

4.1 Datasets

We evaluate our method on three widely used video instance segmentation benchmarks: YouTube-VIS 2019/2021 (YTVIS19/21) and OVIS. The YouTube-VIS datasets contain 40 object categories and serve as standard benchmarks for VIS. YTVIS19 provides 2,238/302/343 videos for training/validation/testing, respectively. YTVIS21 extends YTVIS19 with improved annotations and a slightly revised label set, increasing the scale to 2,985/421/453 videos. OVIS is designed to evaluate VIS methods under challenging conditions, particularly heavy occlusions. It contains 25 categories with 607/140/154 videos for training/validation/

testing. Compared with YouTube-VIS, OVIS videos are longer (around 12 seconds on average) and contain more crowded scenes with higher instance density (about 5.8 instances per video), resulting in frequent severe occlusions and complex object interactions.

4.2 Implementation Details

Our model is built upon Mask2Former [4] with its official hyperparameters unless otherwise specified. All models are initialized with COCO-pretrained weights [13]. Following prior online VIS works [10, 11, 20], we adopt the COCO joint training (CJT) setting. We follow CTVIS [20] for the similarity-guided momentum update [21] and use the same α.

For training, the clip length is set to 8 frames on YTVIS19/21 and 10 frames on OVIS. We apply clip-level random cropping and horizontal flipping as data augmentation. Input frames are resized such that the shorter side is randomly sampled from $[320, 640]$ while the longer side does not exceed 768. During inference, frames are resized to 480p resolution.

The loss weights λ_{gate}, λ_{vpcl}, λ_{cls}, λ_{ce}, and λ_{dice} are set to 5.0, 2.0, 2.0, 5.0, and 5.0, respectively. For the gating regularization loss $\mathcal{L}_{\text{gate}}$, we set the threshold to $\gamma = 0.1$ and the margin to $\mu = 0.01$ on YTVIS19/21, and use $\gamma = 0.05$ and $\mu = 0.005$ on OVIS. Training uses a mini-batch size of 16 for 16,000 iterations. The initial learning rate is 1×10^{-4} and is decayed at 6,000 and 12,000 iterations.

4.3 Main Results

Table 1 compares our method with recent state-of-the-art VIS approaches on YTVIS19/21 [18] and OVIS [15] using two backbones, ResNet-50 [5] and Swin-L [14].

YTVIS19 & YTVIS21. On the YouTube-VIS benchmarks, our method achieves competitive or superior performance across both backbones. With ResNet-50, our method obtains 55.6 AP on YTVIS19 and achieves the best performance of 51.7 AP on YTVIS21. Using the stronger Swin-L backbone, our method further reaches 66.2 AP on YTVIS19 and 61.6 AP on YTVIS21, remaining competitive with recent state-of-the-art methods. These results indicate that the proposed query-level gating, VPCL, and dual-memory association jointly improve the robustness of online instance association. We also improve stricter metrics such as AP_{75} (e.g., 56.6 vs. 54.9 on YTVIS21 with ResNet-50), indicating more precise masks and more stable identity assignment.

OVIS. OVIS is significantly more challenging due to heavy occlusions and crowded scenes, where association errors easily accumulate. Our method achieves the best performance under both backbones, improving the previous best AP from 37.6 to 38.2 with ResNet-50 and from 48.6 to 49.1 with Swin-L compared with CAVIS [11]. The improvements demonstrate that our dual-memory online association together with gated contrastive learning effectively suppresses background interference and stabilizes identity matching under severe occlusions.

Table 2: Ablation study of the three proposed components: query-level gating, VPCL, and dual-memory fusion on YTVIS21 and OVIS.

Query-level gating	VPCL	Dual-memory fusion	YTVIS21 AP	OVIS AP
			47.5	34.7
✓			49.8	35.9
	✓		48.7	36.3
		✓	48.1	35.5
✓	✓		51.4	37.8
✓	✓	✓	**51.7**	**38.2**

Table 3: Ablation study on applying query-level gating during training and/or inference on YTVIS21 and OVIS.

Setting	YTVIS21 AP	OVIS AP
Training-only gating	39.6	25.2
Inference-only gating	48.3	34.3
Training + Inference gating	**51.7**	**38.2**

4.4 Ablation Studies

We conduct ablation studies on YouTubeVIS21 and OVIS following the standard evaluation protocol. Unless otherwise specified, all experiments share the same backbone, training schedule, and inference settings, and only the component under study is modified. These ablations consistently verify the effectiveness of each component in improving embedding quality and association robustness.

Component ablation of MiNQVIS. Table 2 evaluates the three main components of our framework: query-level gating, VPCL, and dual-memory matching. The baseline corresponds to CTVIS [20] with its memory noise injection disabled. Starting from this baseline (47.5/34.7 AP on YTVIS21/OVIS), each component individually improves performance, with the query-level gating providing the largest single-module gain. The query-level gating increases AP to 49.8/35.9 by filtering unreliable queries, while VPCL improves AP to 48.7/36.3 by providing more stable identity supervision. The fused matching module further improves the baseline to 48.1/35.5 by enhancing temporal association.

Combining the gate and prototype leads to a much larger gain (51.4/37.8), indicating strong complementarity between noise-aware query filtering and prototype-based representation learning. Finally, enabling all components achieves the best performance of **51.7/38.2**, demonstrating that improving both embedding quality and matching robustness is critical for stable online VIS.

When to apply the gate. Table 3 studies whether the foreground gate should be applied during training and/or inference. Applying the gate only at infer-

Table 4: Ablation study of dual-memory fusion on YTVIS21 and OVIS.

Tracker memory	YTVIS21 AP	OVIS AP
Momentum memory	51.4	37.8
History-best memory	51.1	36.1
Dual-memory fusion (momentum + history-best)	**51.7**	**38.2**

Table 5: Sensitivity analysis of the gate threshold γ on YTVIS21 and OVIS.

γ	YTVIS21 AP	OVIS AP
0.025	48.5	37.5
0.05	50.3	**38.2**
0.10	**51.7**	36.7
0.20	49.6	32.8

ence already improves performance (48.3/34.3), confirming that filtering low-confidence queries benefits online association. In contrast, using the gate only during training significantly degrades performance (39.6/25.2), suggesting a severe train–test mismatch when the model is trained with gated samples but evaluated without the same mechanism. Applying the gate consistently during both training and inference achieves the best results (51.7/38.2), highlighting the importance of maintaining consistent query selection throughout the pipeline.

Dual-memory matching. Table 4 evaluates the proposed dual-memory association strategy. Using only the momentum memory achieves 51.4/37.8 AP, demonstrating its ability to capture short-term temporal continuity. Using only the history-best memory yields slightly lower performance (51.1/36.1), as relying on a single historical snapshot may not fully adapt to appearance changes. Combining the two memories achieves the best performance (51.7/38.2), confirming that short-term adaptation and reliable historical references are complementary for robust online association.

Sensitivity to the gate threshold. Table 5 analyzes the sensitivity to the gating threshold γ. Very small thresholds retain many noisy queries and limit the improvement (48.5/37.5 at 0.025), while overly large thresholds remove too many valid queries and degrade performance, especially on OVIS (49.6/32.8 at 0.20). The optimal threshold differs across datasets: YTVIS21 achieves the best performance at $\gamma = 0.10$ (51.7 AP), while OVIS prefers a slightly smaller threshold $\gamma = 0.05$ (38.2 AP), reflecting the need to retain more candidate queries under severe occlusion.

4.5 Qualitative Results

Qualitative comparison. Fig. 3 shows qualitative comparisons under heavy occlusion, where multiple similar instances undergo complex motion and fre-

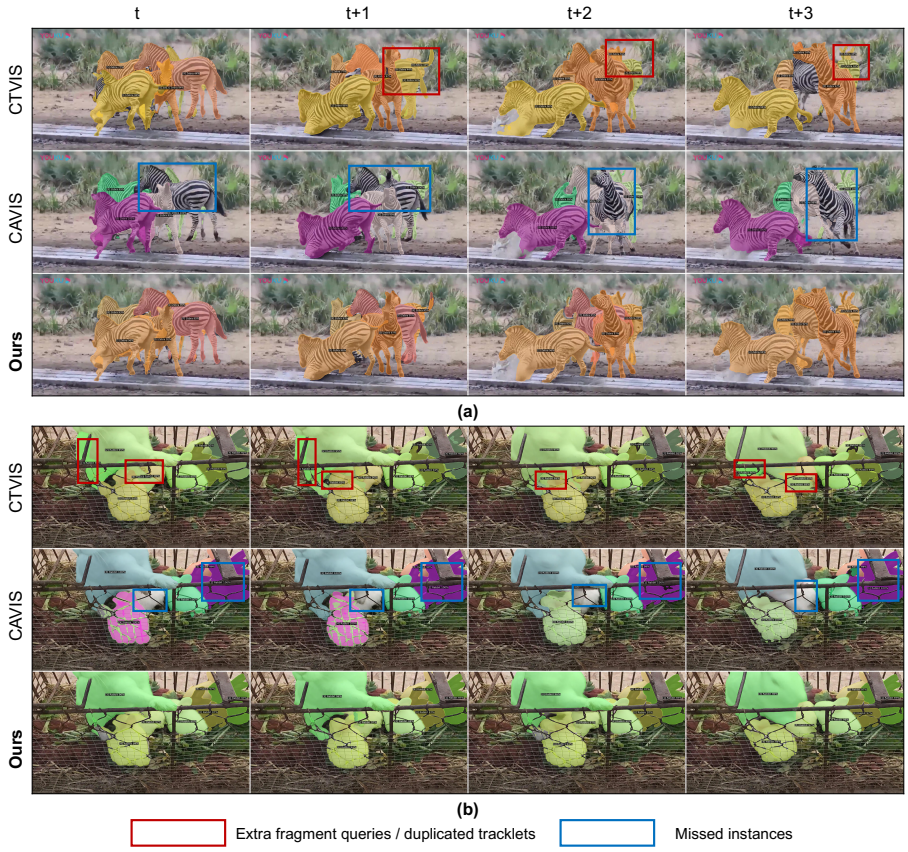

Fig. 3: Qualitative comparison under occlusion-induced fragmentation. We compare CTVIS, CAVIS, and our method on two challenging clips from **OVIS**, where heavy occlusion causes masks to become fragmented across frames. Each row corresponds to a method and columns show consecutive frames. Our method maintains cleaner masks and more consistent identities across frames.

quent mutual blocking, often causing fragmented masks and unstable associations across frames. Such crowded scenarios are particularly challenging for online video instance segmentation, as partial visibility and overlapping objects make it difficult to maintain consistent instance identities over time. CTVIS [20] tends to produce redundant fragmented predictions in heavily occluded regions, which often results in duplicated tracklets and unstable identity assignments across consecutive frames. CAVIS [11] is more conservative in generating predictions, but it frequently misses partially visible objects or fails to associate fragmented regions consistently, leading to missed detections and identity fragmentation. In contrast, our method produces cleaner and more coherent masks even under strong occlusion, and is able to maintain more consistent instance identities across frames. These results indicate that our approach is more ro-

bust to occlusion-induced fragmentation and better preserves temporal identity consistency in crowded and challenging scenes.

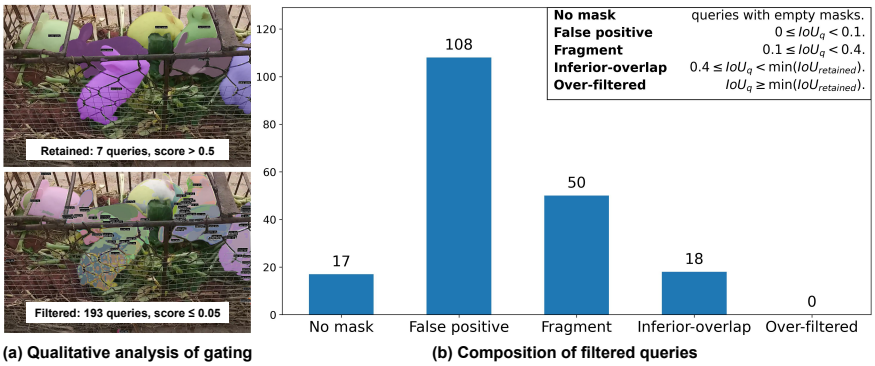

Fig. 4: Qualitative analysis of the gating mechanism. (a) Visualization of the retained and filtered queries on an OVIS example. The top image shows the queries retained after gating, while the bottom image shows the queries filtered out by our gating mechanism. (b) The composition of filtered queries in (a), where IoU_q denotes the query-to-GT IoU. Most filtered queries correspond to false positives or fragmented masks, while no over-filtered query is observed.

Qualitative analysis of gating. Fig. 4 further illustrates how the proposed gating mechanism improves query reliability in crowded scenes. By suppressing unreliable queries before online association, the model avoids introducing noisy masks into track states and reduces redundant or fragmented predictions. This leads to a more compact and reliable query set, which helps maintain cleaner instance masks and more stable identity consistency under heavy occlusion.

Failure cases. Fig. 5 shows representative failure cases. Although our method improves robustness under occlusion and clutter, it can still fail in scenarios with extreme motion or drastic pose changes. In such situations, instance appearance may change significantly across frames, making different objects visually similar or causing partial observations to resemble other instances. This ambiguity can lead to incorrect associations, where an object temporarily inherits the identity of a nearby instance with a similar appearance. These failures typically occur when temporal cues are weak and appearance information becomes unreliable due to rapid motion or severe pose variation.

5 Conclusion

In this work, we revisit online video instance segmentation from a noise-aware perspective and identify noisy instance queries as a key source of instability in

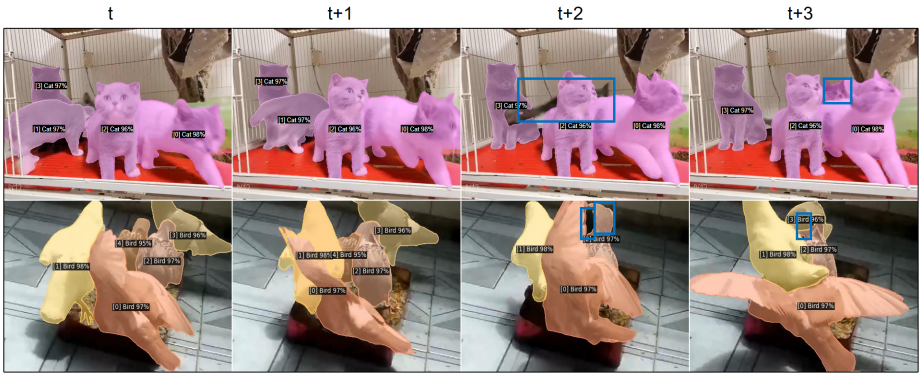

Fig. 5: Failure cases under extreme motion and pose ambiguity. Rapid motion and drastic pose changes can make instance appearance highly ambiguous, occasionally leading to identity mismatches even for our method. In such situations, partial observations or sudden appearance changes may cause different instances to become visually similar, making reliable association difficult. **Blue boxes highlight regions with missed instances or temporary identity confusion.**

temporal association and contrastive supervision. To address this issue, we propose MiNQVIS, a unified framework that improves the robustness of online VIS by stabilizing query representations and association decisions over time. Specifically, we introduce three complementary components: (1) a query-level gating mechanism that filters unreliable predictions via classification-guided thresholding; (2) a video-level prototypical contrastive learning (VPCL) strategy that constructs stable instance prototypes from reliable cross-frame matches; and (3) a dual-memory fusion module that combines momentum embeddings with history-best retrieval to mitigate long-term drift. Together, these components improve both embedding quality and temporal matching reliability in challenging scenarios such as occlusion, motion blur, and crowded scenes. Experiments on YTVIS 2019/2021 and OVIS show consistent improvements over strong online VIS baselines with both ResNet-50 and Swin-L backbones. Ablation studies further validate the effectiveness of each component and highlight the importance of noise-aware representation learning for robust online VIS. Beyond the specific design choices, our results suggest that explicitly modeling query reliability and stabilizing temporal representations are promising directions for improving online video perception systems. We hope this work encourages future research on noise-aware representation learning and more reliable temporal association mechanisms for real-world video understanding.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (No. 62276071), the Guangdong Special Support Program for Science and Technology Innovation Talents (No. 0620220211), and the Guangzhou Key Research and Development Program (No. 2025B01J3011).

References

1. Cao, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., Shao, L.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: European Conference on Computer Vision. pp. 1–18. Springer (2020)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
4. 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. pp. 1290–1299 (2022)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Heo, M., Hwang, S., Hyun, J., Kim, H., Oh, S.W., Lee, J.Y., Kim, S.J.: A generalized framework for video instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14623–14632 (2023)
7. Heo, M., Hwang, S., Oh, S.W., Lee, J.Y., Kim, S.J.: Vita: Video instance segmentation via object token association. Advances in neural information processing systems **35**, 23109–23120 (2022)
8. Huang, D.A., Yu, Z., Anandkumar, A.: Minvis: A minimal video instance segmentation framework without video-based training. Advances in Neural Information Processing Systems **35**, 31265–31277 (2022)
9. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers. Advances in Neural Information Processing Systems **34**, 13352–13363 (2021)
10. Kim, H., Kang, J., Heo, M., Hwang, S., Oh, S.W., Kim, S.J.: Visage: Video instance segmentation with appearance-guided enhancement. In: European Conference on Computer Vision. pp. 93–109. Springer (2024)
11. Lee, S., Seo, J., Han, K., Choi, M., Im, S.: Cavis: Context-aware video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4507–4517 (2025)
12. Li, J., Yu, B., Rao, Y., Zhou, J., Lu, J.: Tcovis: Temporally consistent online video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1097–1107 (2023)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
15. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H., Bai, S.: Occluded video instance segmentation: A benchmark. International Journal of Computer Vision **130**(8), 2022–2039 (2022)

16. Wu, J., Jiang, Y., Bai, S., Zhang, W., Bai, X.: Seqformer: Sequential transformer for video instance segmentation. In: European Conference on Computer Vision. pp. 553–569. Springer (2022)
17. Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. In: European Conference on Computer Vision. pp. 588–605. Springer (2022)
18. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5188–5197 (2019)
19. Yang, S., Fang, Y., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Crossover learning for fast online video instance segmentation. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 8043–8052 (2021)
20. Ying, K., Zhong, Q., Mao, W., Wang, Z., Chen, H., Wu, L.Y., Liu, Y., Fan, C., Zhuge, Y., Shen, C.: Ctvis: Consistent training for online video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 899–908 (2023)
21. Yu, E., Li, Z., Han, S.: Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8834–8843 (2022)
22. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: European conference on computer vision. pp. 659–675. Springer (2022)
23. 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. pp. 1282–1291 (2023)
24. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)