CROSS GRANULARITY DECOUPLED DECODER AND TEMPORAL QUERY EVOLUTION FOR REFERRING VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring Video Object Segmentation (RVOS) aims to segment the target specified by a language expression throughout a video. Recent diffusion based RVOS methods benefit from text aligned spatiotemporal representations, but converting diffusion features into accurate and temporally stable referring masks remains challenging. Frame wise query decoding may suffer from target drift under occlusion, disappearance, and similar distractors. Meanwhile, a unified mask prediction head has to handle both target localization and complete mask generation in the same prediction space. To address these limitations, we propose a prediction stage adaptation framework for diffusion based RVOS. First, we introduce Temporal Query Evolution (TQE), which updates object queries with short term temporal context and text aligned long range references to maintain target identity over time. Second, we design a Cross Granularity Decoupled Decoder (CGD), which decomposes mask prediction into discriminative localization and semantic coverage. The Discriminative Instance Branch performs query conditioned instance filtering to identify the referred target, while the Holistic Semantic Branch provides broader semantic activation for mask completion. We employ Temporal Context Mask Refinement (TCMR) to refine low resolution masks with temporal boundary cues and improve high resolution mask quality. Experiments on five RVOS benchmarks validate the framework's effectiveness and generalization.

1 INTRODUCTION

Referring Video Object Segmentation (RVOS) Wang et al. (2026); Liang et al. (2025); Wu et al. (2022b) aims to segment the object described by a language expression throughout a video. The model must identify the referred target, distinguish it from similar objects, and keep its identity across frames. Occlusion, appearance changes, scale variation, and distractors pose challenges.

ReferEverything Bagchi et al. (2025) and FlowRVS Wang et al. (2026) adapt generative models or VAE decoders to produce segmentation masks, while VD-IT Zhu et al. (2024) uses a pretrained text-to-video diffusion model as a feature backbone. However, generative features are not directly optimized for discriminative mask prediction. Thus, the prediction stage remains important for converting strong video-language features into accurate, complete, and temporally stable masks.

In the VD-IT-style pipeline, masks are predicted from frame-wise queries with a unified mask head. First, the query is guided by the same language expression in each frame, while the target appearance changes over time. It may drift under occlusion, disappearance, or similar distractors. Second, the unified head mixes target selection and mask completion in one prediction space. Target selection requires instance-level discrimination, while mask completion requires broader semantic coverage.

To address these issues, we revisit the prediction stage from query evolution, mask prediction, and mask refinement. TQE updates object queries with short-term temporal context and text-aligned long-range references to maintain target identity. CGD decomposes mask prediction into a Holistic Semantic Branch for region coverage and a Discriminative Instance Branch for target-specific localization, better supporting multi-object coverage and single-object discrimination. TCMR further uses temporal context during mask refinement to improve mask quality and temporal coherence.

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

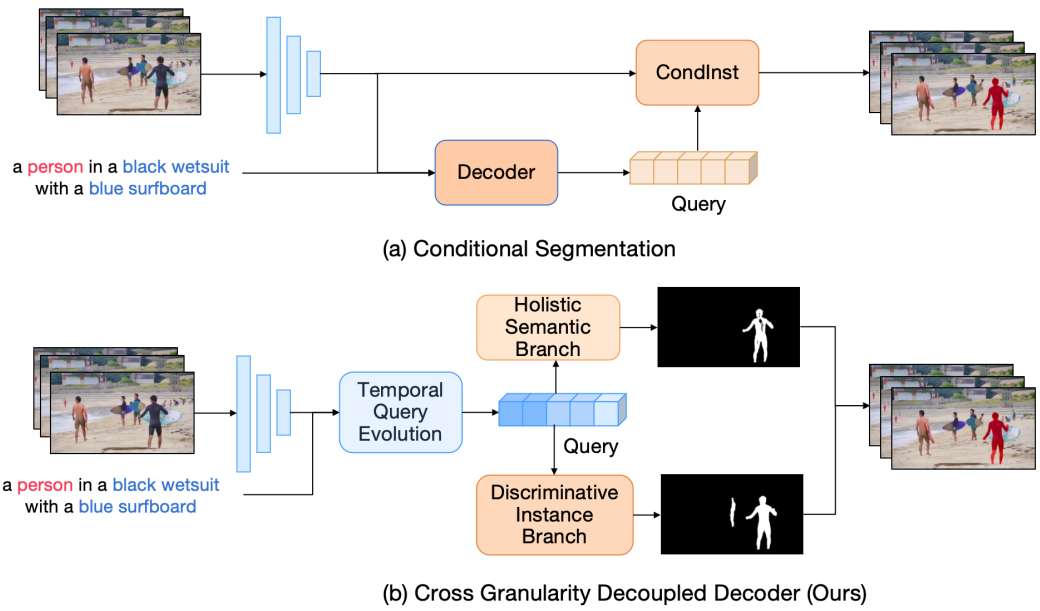

Figure 1: Unified mask prediction versus CGD. The unified head uses a single prediction space, while CGD separates target localization and mask coverage into two complementary branches.

Our contributions can be summarized as follows:

- We design a Cross Granularity Decoupled Decoder to reduce tension between segmentation needs. It splits mask prediction into a Holistic Semantic Branch and a Discriminative Instance Branch, which support multi object coverage and single object discrimination.
- We propose Temporal Query Evolution for object queries. It combines local temporal interaction with text aligned long range references to preserve query continuity.
- We introduce a Temporal Context Mask Refinement module to capture inter frame dependencies during refinement. It improves quality and temporal coherence with minimal cost.
- Experiments on five standard RVOS benchmarks validate the effectiveness of the proposed modules and demonstrate consistent improvements under the controlled VD-IT pipeline.

2 Related Work

2.1 Referring Video Object Segmentation

URVOS (Seo et al., 2020) introduced cross-modal attention and memory attention in a unified framework. MLRL Wu et al. (2022a) learned multi-level visual features and aligned them with language.

Later, transformer based methods improved video and language modeling. MTTR (Botach et al., 2022) formulated RVOS as multimodal sequence prediction. ReferFormer Wu et al. (2022b) used object queries guided by language for mask prediction. SgMg Miao et al. (2023) improved visual text alignment with spectrum guided fusion and refined masks with a multi granularity optimizer.

SAMWISE Cuttano et al. (2025) adapts SAM2 for text-driven video segmentation with temporal and multimodal adapters. VD-IT Zhu et al. (2024) uses a pretrained text-to-video diffusion model as a feature backbone. ReferEverything Bagchi et al. (2025) extends video diffusion models to open-vocabulary referring video segmentation with a VAE-based mask representation. FlowRVS Wang et al. (2026) formulates RVOS as a conditional continuous flow and adapts pretrained text-to-video generative models Team Wan et al. (2025) for mask generation. These methods show that generative video models provide semantic and temporal priors. However, generating masks from generative representations often needs task-specific adaptation of the decoder or VAE. This suggests an objective gap between generative reconstruction and discriminative segmentation.

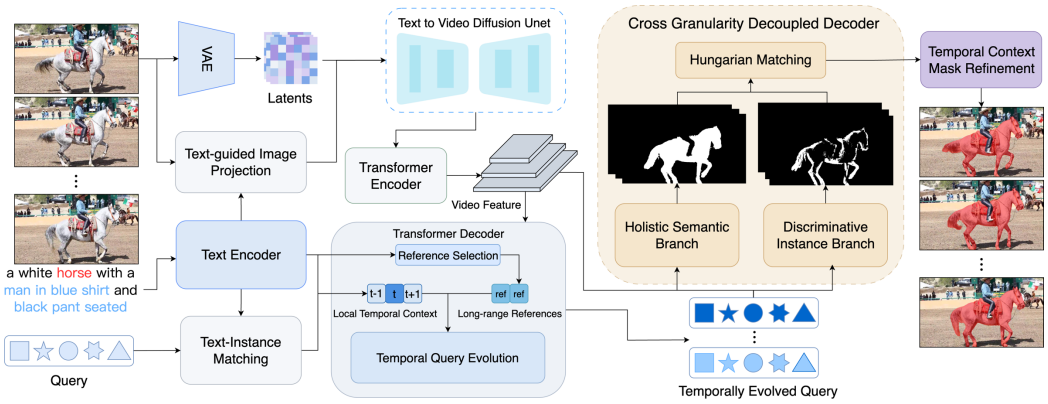

Figure 2: Framework overview. We extract diffusion based video features and initialize object queries from language features. TQE updates the queries with local temporal context and text aligned long range references. CGD then decomposes mask prediction into discriminative localization and semantic coverage. TCMR refines low resolution masks with temporal boundary cues.

We study prediction stage adaptation on a diffusion based RVOS backbone instead of adapting a generative decoder or VAE. Our method decouples mask prediction into discriminative localization and semantic coverage, with temporal adaptation at the object query and mask refinement levels.

2.2 Temporal Reasoning and Propagation

Temporal modeling is central to RVOS because the target may deform, disappear, or be occluded. URVOS Seo et al. (2020) uses memory attention to improve temporal consistency. OnlineRefer Wu et al. (2023) propagates object queries online. FindTrack Cho et al. (2025) separates target identification from mask propagation by selecting a reliable key frame. LoSh Yuan et al. (2024) uses optical flow to align features between annotated frames and nearby frames during training. ReferDINO Liang et al. (2025) combines a memory-augmented tracker with a cross-modal temporal decoder for object-consistent temporal reasoning. In contrast, TQE focuses on query adaptation at the prediction stage. It updates the current object query with neighboring context and text aligned long range references. This improves target modeling without heavy temporal propagation.

3 Method

3.1 Overall Pipeline

We adapt VD-IT Zhu et al. (2024) at the prediction stage for diffusion-based RVOS. Given a video clip $V = \{I_t\}_{t=1}^{T}$ and a referring expression X, the goal is to predict a target mask $\hat{M}^t$ for each frame I_t. Figure 2 shows three components: Temporal Query Evolution (TQE), Cross Granularity Decoupled Decoder (CGD), and Temporal Context Mask Refinement (TCMR). Following VD-IT, the input video is encoded by a VAE Kingma & Welling (2014) into latent representations, and the expression is encoded by CLIP Radford et al. (2021) for text conditioned diffusion feature extraction. A pretrained text to video diffusion backbone fuses the video latents and text features. The Transformer Encoder produces multi level features $F_v^t = \{F_{\times 8}^t, F_{\times 16}^t, F_{\times 32}^t\}$. In parallel, RoBERTa Liu et al. (2019) encodes the referring expression to initialize object queries, which are processed by a Deformable Transformer Decoder Zhu et al. (2021). Our modules follow feature extraction.

3.2 Temporal Query Evolution

We propose Temporal Query Evolution (TQE) to address query drift in diffusion-based RVOS decoding. In existing frame-wise decoding, object queries are updated mainly within each frame, so the query assigned to the referred target may lose identity cues under occlusion, disappearance, or similar distractors. A direct solution is to apply dense temporal attention over all frames, but this may introduce noisy query interactions from irrelevant frames. TQE instead performs text-guided

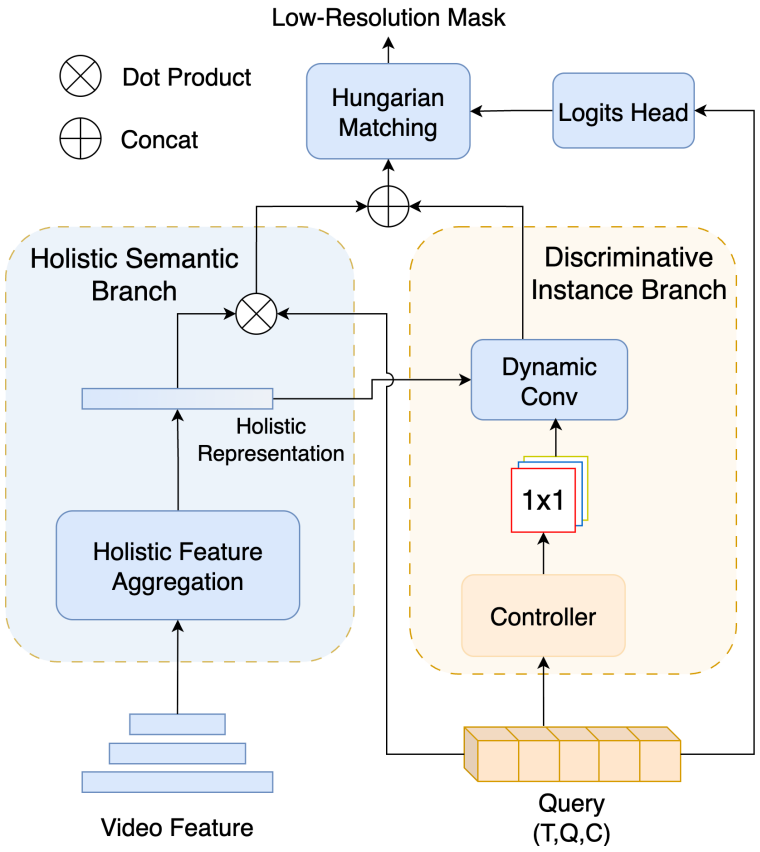

Figure 3: CGD uses HSB for mask completion and DIB for query-conditioned target localization.

sparse query evolution, where each query is updated only with temporally reliable and text-relevant evidence. Given the decoder queries of a clip, we reshape them from $[BT, Q, C]$ to $[T, BQ, C]$ for temporal interaction. TQE builds two types of temporal evidence. The first type is local evidence from adjacent frames, which preserves short-term query continuity. The second type is long-range evidence selected according to the referring expression, which helps recover target cues when the local neighborhood is unreliable. To obtain the long-range evidence, we compute a global text feature E_x by averaging word features, and represent each frame by the average of its object queries. The relevance score of frame i is computed as

$$s_i = \cos(\phi_q(Q_i), \phi_x(E_x)), \tag{1}$$

where $\phi_q(\cdot)$ and $\phi_x(\cdot)$ are projection layers. We select the top-k frames with the highest scores as text-aligned long-range references, where $k = 3$. Based on these references, we build a sparse temporal attention mask. For each frame, the query can attend to its adjacent frames and the selected text-aligned long-range frames, while other frames are masked out. We further inject rotary temporal positional embeddings into the query and key of temporal attention to encode frame order.

3.3 Cross Granularity Decoupled Decoder

A unified mask head predicts masks in a single prediction space. This forces target selection and mask completion to share the same representation. Target localization needs instance discrimination, whereas mask completion needs broad semantic coverage. CGD therefore decomposes mask prediction into a Holistic Semantic Branch (HSB) and a Discriminative Instance Branch (DIB).

As shown in Fig. 3, given the decoder query Q_t at frame t, we build a shared visual feature from encoder features with downsampling ratios of $8\times$, $16\times$, and $32\times$:

$$F_m^t = \text{Conv}\left(\text{Concat}\left(F_{\times 8}^t, F_{\times 16}^t, F_{\times 32}^t\right)\right). \tag{2}$$

We upsample features to a common size before fusing spatial and semantic cues for both branches.

HSB uses DETR-style dot products (Carion et al., 2020) for semantic coverage:

$$M_h^t = \phi_q^h(Q_t)\phi_f^h(F_m^t)^\top, \tag{3}$$

where $\phi_q^h(\cdot)$ and $\phi_f^h(\cdot)$ are query and feature projections. The resulting semantic activation helps recover complete object regions. DIB uses CondInst-style dynamic convolution (Tian et al., 2020):

$$M_d^t = \phi_q^d(Q_t) * \phi_f^d(F_m^t), \tag{4}$$

where $\phi_q^d(\cdot)$ generates an instance-specific kernel, $\phi_f^d(\cdot)$ projects the visual feature, and $*$ denotes dynamic convolution. This filtering distinguishes the target from similar objects. We fuse the branch outputs for mask prediction and use the evolved query to predict confidence and bounding boxes:

$$\begin{aligned} M_l^t &= P_m\left(\mathrm{Concat}\left(M_h^t, M_d^t\right)\right), \\ \left(S^t, B^t\right) &= \left(P_s(Q_t), P_b(Q_t)\right). \end{aligned} \tag{5}$$

Here, M_l^t is the low-resolution mask, S^t the confidence score, and B^t the predicted bounding box.

During training, Hungarian matching assigns one prediction to the ground-truth target:

$$j^* = \arg\min_j \mathcal{C}_{\mathrm{match}}\left(M_{l,j}^t, S_j^t, B_j^t, Y^t\right), \tag{6}$$

where $\mathcal{C}_{\mathrm{match}}$ considers mask quality, confidence, and bounding-box consistency. The matched mask is used for supervision and subsequent refinement to improve final mask quality. During inference, we select the query with the highest confidence averaged across all video frames.

3.4 Temporal Context Mask Refinement

Temporal Context Mask Refinement (TCMR) recovers high-resolution masks with sharper, more stable boundaries from M_l^t. M_l^t locates the target, but direct upsampling produces coarse contours and frame-wise detail changes. This problem is more obvious for fast motion, occlusion, and small objects, where local boundaries are hard to recover from low-resolution diffusion features alone.

TCMR refines the mask in two stages. First, it uses the $8\times$ diffusion feature $F_{\times 8}^t$ to restore the coarse target structure. Second, it uses a motion-aware $4\times$ feature to refine local boundaries. To build this feature, the raw frames in the input clip are downsampled by two stride-2 convolution layers and a 1×1 projection. A temporal convolution is then applied over the clip, and its output is added back to the spatial feature to produce $F_{\times 4}^t$. This design brings short-term motion cues into high-resolution mask refinement without modifying the diffusion backbone. The refinement process is defined as

$$\hat{M}^t = \mathrm{Up}_{\times 4}\Big(\mathrm{Conv}\big(\mathrm{Up}_{\times 2}(\mathrm{Conv}(M_l^t \oplus F_{\times 8}^t)) \oplus F_{\times 4}^t\big)\Big). \tag{7}$$

where $\oplus$ denotes channel-wise concatenation.

3.5 Loss Function

The loss combines mask prediction, confidence scoring, and bounding box regression. For mask prediction, we combine Dice Loss Milletari et al. (2016) and Focal Loss Lin et al. (2017), and apply them to the final mask output and the intermediate low-resolution mask. The mask loss is defined as

$$\begin{aligned} \mathcal{L}_{mask} =& \lambda_{focal}\mathcal{L}_{mask_focal} + \lambda_{dice}\mathcal{L}_{mask_dice} \\ &+ \lambda_{focal}\mathcal{L}_{mask_low_focal} + \lambda_{dice}\mathcal{L}_{mask_low_dice}. \end{aligned} \tag{8}$$

Confidence uses Focal Loss; box regression uses L_1 and GIoU (Rezatofighi et al., 2019).

The Transformer decoder has four layers. Following DETR, we compute the main loss on the output of the final layer, while the outputs of the first three layers are used as auxiliary supervision:

$$\mathcal{L}_{aux} = \sum_{l=1}^{3}\left[\mathcal{L}_{mask}^{(l)} + \lambda_{box}\mathcal{L}_{box}^{(l)} + \lambda_{logits}\mathcal{L}_{logits}^{(l)}\right]. \tag{9}$$

The total training loss is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{mask} + \lambda_{box}\mathcal{L}_{box} + \lambda_{logits}\mathcal{L}_{logits} + \mathcal{L}_{aux}. \tag{10}$$

Here, $\lambda_{(\cdot)}$ denotes the balancing coefficient. $\mathcal{L}_{mask}$ denotes the mask segmentation loss, $\mathcal{L}_{box}$ denotes the bounding box regression loss, and $\mathcal{L}_{logits}$ denotes the confidence score loss.

Method	Publication	MeViS			Ref-YouTube-VOS			Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Methods with LLMs										
VideoLISA-3.8B Bai et al. (2024)	NeurIPS'24	44.4	41.3	47.6	63.7	61.7	65.7	68.8	64.9	72.7
VISA-13B Yan et al. (2024)	ECCV'24	44.5	41.8	47.1	63.0	61.4	64.7	70.4	67.0	73.8
GLUS-7B Lin et al. (2025)	CVPR'25	51.3	48.5	54.2	67.3	65.5	69.0	–	–	–
RGA3-7B Wang et al. (2025)	ICCV'25	50.1	47.4	52.8	68.5	66.8	70.1	72.8	68.3	77.3
VRS-HQ-13B Gong et al. (2025)	CVPR'25	50.9	48.0	53.7	71.0	69.0	73.1	74.4	71.0	77.9
Methods without LLMs										
ReferFormer Wu et al. (2022b)	CVPR'22	31.0	29.8	32.2	62.9	61.3	64.6	61.1	58.1	64.1
MLRL Wu et al. (2022a)	CVPR'22	–	–	–	49.7	48.4	51.0	52.8	50.0	55.4
SgMg Miao et al. (2023)	ICCV'23	–	–	–	65.7	63.9	67.4	63.3	60.6	66.0
VD-IT Zhu et al. (2024)	ECCV'24	–	–	–	66.5	64.4	68.5	69.4	66.2	72.6
ReferDINO Liang et al. (2025)	ICCV'25	49.3	44.7	53.9	69.3	67.0	71.5	68.9	65.1	72.9
DMVS Fang et al. (2025)	CVPR'25	48.6	44.2	52.9	64.3	62.4	66.2	65.2	62.2	68.2
SAMWISE Cuttano et al. (2025)	CVPR'25	49.5	46.6	52.4	69.2	67.8	70.6	70.6	67.4	74.5
FlowRVS Wang et al. (2026)	ICLR'26	51.1	47.6	54.6	69.6	67.1	72.1	73.3	68.4	**78.2**
CGD-2.6B	**Ours**	**54.9**	**51.8**	**58.0**	**72.0**	**69.9**	**74.1**	**74.6**	**72.1**	77.2

Table 1: Results on MeViS, Ref-YouTube-VOS, and Ref-DAVIS17 validation sets.

Method	Publication	A2D-Sentences			JHMDB-Sentences		
		mAP	oIoU	mIoU	mAP	oIoU	mIoU
ReferFormer Wu et al. (2022b)	CVPR'22	52.8	77.6	69.6	42.2	71.9	71.0
SgMg Miao et al. (2023)	ICCV'23	56.1	78.0	70.4	44.4	72.8	71.7
VD-IT Zhu et al. (2024)	ECCV'24	61.4	81.5	73.2	46.5	74.4	73.4
LoSh Yuan et al. (2024)	CVPR'24	59.9	81.2	73.1	-	-	-
DsHmp He & Ding (2024)	CVPR'24	59.8	81.1	72.9	45.8	73.9	73.0
ReferDINO Liang et al. (2025)	ICCV'25	61.1	**82.1**	73.6	46.6	74.2	73.2
CGD	**Ours**	**62.5**	82.0	**74.2**	**47.1**	**75.2**	**73.9**

Table 2: Comparison on the A2D-Sentences and JHMDB-Sentences validation sets.

4 Experiment

4.1 Datasets and Evaluation Metrics

We compare methods under their reported training settings. Since training data and pretraining strategies vary across methods, Tables 1 and 2 evaluate system performance, while Table 3 provides a controlled comparison under the same training pipeline. We evaluate our method on five benchmarks, including Ref-YouTube-VOS Seo et al. (2020), Ref-DAVIS17 Khoreva et al. (2018), MeViS Ding et al. (2023), A2D Sentences Gavrilyuk et al. (2018), and JHMDB Sentences Gavrilyuk et al. (2018). For Ref-YouTube-VOS, Ref-DAVIS17, and MeViS, we follow the standard evaluation protocol and report region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$. $\mathcal{J}$ measures the overlap between the predicted mask and the ground-truth mask, while $\mathcal{F}$ evaluates boundary quality. For A2D Sentences and JHMDB Sentences, we report mAP, overall IoU, and mean IoU.

4.2 Implementation Details

We build our method on the pretrained text-to-video diffusion model ModelScopeT2V Wang et al. (2023). The VAE, CLIP, RoBERTa, and diffusion backbone are kept frozen during training. We apply LoRA to the backbone Transformer's linear layers, updating only the query and value projections. The LoRA rank and alpha are set to 16 and 32, respectively. We use a Deformable Transformer encoder and decoder. The model has 2.6B parameters, of which 3M are trainable. We use 5 instance queries. We use offline, clip-level inference. On Ref-YouTube-VOS, it runs at 19 FPS on one H20 GPU.

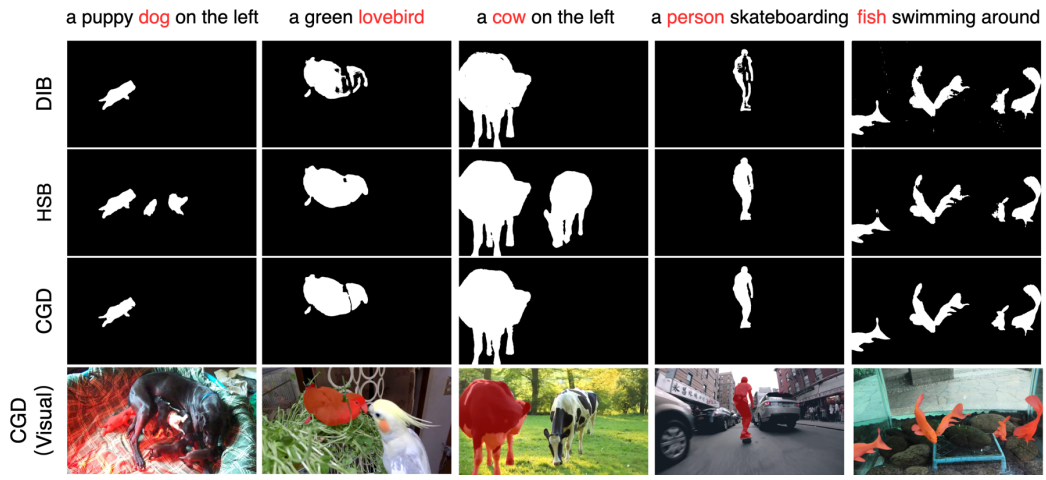

Figure 4: Branch comparison. DIB focuses on the referred target, while HSB provides broader mask responses that may include several instances. CGD combines both for more accurate masks.

Pretrain	Finetune	Head	J&F	J	F
None	Ref-YT-VOS	Unified	70.46	68.33	72.59
MeViS	Ref-YT-VOS	Unified	69.23	67.61	70.85
None	Ref-YT-VOS	Decoupled	71.10	68.92	73.28
MeViS	Ref-YT-VOS	Decoupled	**72.03**	**69.96**	**74.10**

Table 3: Effect of pretraining and segmentation head design.

We pretrain on MeViS and fine-tune on the mixed Ref-YouTube-VOS and Ref-DAVIS17 training set. During pretraining, the model is trained on 8 H20 GPUs for 10 epochs, taking about 6 days. During fine-tuning, it is trained on the same hardware for 6 epochs, taking about 20 hours. A2D Sentences and JHMDB Sentences are used only for evaluation, not training or fine-tuning.

4.3 Comparison with State-of-the-Art Methods

MeViS. On MeViS, CGD achieves 54.9 in $\mathcal{J}\&\mathcal{F}$, 51.8 in $\mathcal{J}$, and 58.0 in $\mathcal{F}$. CGD leads the compared methods on all three metrics, improving $\mathcal{J}\&\mathcal{F}$ by 3.6 points over the next-highest result.

Ref-YouTube-VOS. On Ref-YouTube-VOS, CGD achieves 72.0 in $\mathcal{J}\&\mathcal{F}$, 69.9 in $\mathcal{J}$, and 74.1 in $\mathcal{F}$. This leads the compared methods in $\mathcal{J}\&\mathcal{F}$, exceeding diffusion-based FlowRVS by 2.4 points.

Ref-DAVIS17. On Ref-DAVIS17, CGD achieves 74.6 in $\mathcal{J}\&\mathcal{F}$, 72.1 in $\mathcal{J}$, and 77.2 in $\mathcal{F}$. CGD achieves the highest $\mathcal{J}\&\mathcal{F}$ and $\mathcal{J}$ scores among the compared methods, with a competitive $\mathcal{F}$ score.

A2D-Sentences and JHMDB-Sentences. On A2D-Sentences, it obtains 62.5 mAP and 74.2 mIoU, leading the compared methods, with oIoU only 0.1 below the best. On JHMDB-Sentences, CGD leads all compared methods on all three metrics, with 47.1 mAP, 75.2 oIoU, and 73.9 mIoU.

4.4 Ablation Studies

Cross Granularity Conflict Analysis.

Table 3 compares segmentation heads under multi-object pretraining. After pretraining on MeViS and fine-tuning on Ref-YouTube-VOS, the unified head drops from 70.46 to 69.23 in $\mathcal{J}\&\mathcal{F}$. This suggests that a unified prediction head does not adapt well when the model is transferred from multi-object scenarios to a single-target task. With the same training pipeline, MeViS pretraining improves the performance of the decoupled design from 71.10 to 72.03 in $\mathcal{J}\&\mathcal{F}$. Without pretraining, the decoupled head outperforms the unified head by 0.64 in $\mathcal{J}\&\mathcal{F}$. After MeViS pretraining, the margin increases to 2.80. The decoupled head thus benefits more from multi-object pretraining.

Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CGD	**72.03**	**69.96**	**74.10**
w/o DIB	69.74	67.53	71.94
w/o HSB	70.48	68.36	72.60

Table 4: Ablation study of DIB and HSB on Ref-YouTube-VOS.

Method		**Score**		
TQE	Language-Initialized Query	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✓	✓	**72.03**	**69.96**	**74.10**
✓	✗	71.44	69.41	73.48
✗	✗	71.28	69.36	73.20
✗	✓	70.70	68.82	72.58

Table 5: Ablation study of TQE and language-initialized query.

Cross Granularity Decoupled Decoder Analysis.

Table 4 reports the ablation results of DIB and HSB on Ref-YouTube-VOS. All variants are retrained under the same setting for fair comparison. Removing DIB decreases $\mathcal{J}\&\mathcal{F}$ from 72.03 to 69.74, with drops of 2.43 in $\mathcal{J}$ and 2.16 in $\mathcal{F}$. This shows that DIB contributes to target localization and instance-level discrimination. Removing HSB decreases $\mathcal{J}\&\mathcal{F}$ to 70.48, with consistent drops in both $\mathcal{J}$ and $\mathcal{F}$. Removing HSB weakens semantic coverage and mask completeness.

Temporal Query Evolution Analysis.

Table 5 shows that language initialization without TQE gives the lowest $\mathcal{J}\&\mathcal{F}$ score of 70.70. A fixed language-guided query is less effective under temporal changes in the video.

Using TQE without language initialization improves $\mathcal{J}\&\mathcal{F}$ from 71.28 to 71.44. Language initialization alone can therefore hurt performance without temporal adaptation. The best result is achieved when language initialization and TQE are both used, reaching 72.03 in $\mathcal{J}\&\mathcal{F}$.

Mask Refinement and Noise Prediction Analysis.

Table 6 compares different refinement methods and the noise prediction module. Replacing VAE with CNN brings only marginal improvement. Specifically, the $\mathcal{J}\&\mathcal{F}$ score increases from 70.83 to 71.06 when noise prediction is enabled, and from 71.01 to 71.22 when noise prediction is removed.

Replacing CNN with TCMR improves $\mathcal{J}\&\mathcal{F}$ from 71.06 to 71.69 with noise prediction, and from 71.22 to 72.03 without it. TCMR improves $\mathcal{F}$ more than $\mathcal{J}$, suggesting better boundary recovery.

Removing noise prediction improves all refinement variants in Table 6. Noise prediction appears less suited to discriminative mask refinement. We therefore use TCMR without noise prediction.

4.5 Qualitative Comparison

We compare the qualitative results of our method with VD-IT in Fig. 5. In the first example, several zebras with similar appearance appear in the same scene. VD-IT gradually drifts to nearby distractors, while our method keeps more stable predictions and better preserves the target identity. In the second example, the referred target is small and changes its appearance over time. VD-IT produces incomplete, unstable masks, while our method gives stable localization and cleaner masks.

Mask Refinement	NP	Performance		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
VAE	✓	70.83	68.98	72.68
VAE	✗	71.01	69.15	72.88
CNN	✓	71.06	69.13	73.00
CNN	✗	71.22	69.39	73.05
TCMR	✓	71.69	69.71	73.66
TCMR	✗	**72.03**	**69.96**	**74.10**

Table 6: Ablation of mask refinement and noise prediction (NP).



Figure 5: Qualitative comparison of CGD and VD-IT on challenging videos. Our method shows better target localization, mask completeness, and temporal stability across frames.

5 Conclusion

We study prediction stage design for diffusion-based RVOS. Temporal Query Evolution improves target modeling over time, the Cross Granularity Decoupled Decoder separates target discrimination from mask completion, and Temporal Context Mask Refinement sharpens masks using temporal cues. On five benchmarks, these components improve segmentation accuracy and temporal stability.

REFERENCES

Anurag Bagchi, Zhipeng Bao, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Refer-Everything: Towards segmenting everything we can speak of in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23221–23231, 2025.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6833–6859, 2024.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4985–4995, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Suhwan Cho, Seunghoon Lee, Minhyeok Lee, Jungho Lee, and Sangyoun Lee. Find first, track next: Decoupling identification and propagation in referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3873–3883, 2025.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAM-WISE: Infusing wisdom in SAM2 for text-driven video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3395–3405, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2694–2703, 2023.

Hao Fang, Runmin Cong, Xiankai Lu, Xiaofei Zhou, Sam Kwong, and Wei Zhang. Decoupled motion expression video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13821–13831, 2025.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5958–5966, 2018.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29183–29192, 2025.

Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13332–13341, 2024.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pp. 123–141. Springer, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. ReferDINO: Referring video object segmentation with visual grounding foundations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20009–20019, 2025.

Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. GLUS: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8658–8667, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 920–930, 2023.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision*, pp. 565–571. IEEE, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pp. 208–223. Springer, 2020.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision*, pp. 282–298. Springer, 2020.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22274–22284, 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Zanyi Wang, Dengyang Jiang, Liuzhuozheng Li, Sizhe Dang, Chengzu Li, Harry Yang, Guang Dai, Mengmeng Wang, and Jingdong Wang. Deforming videos to masks: Flow matching for referring video segmentation. In *International Conference on Learning Representations*, 2026.

Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-level representation learning with semantic alignment for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4996–5005, 2022a.

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2761–2770, 2023.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4974–4984, 2022b.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pp. 98–115. Springer, 2024.

Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. LoSh: Long-short text joint prediction network for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14001–14010, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In *European Conference on Computer Vision*, pp. 452–469. Springer, 2024.

APPENDIX

A ADDITIONAL IMPLEMENTATION DETAILS

A.1 ARCHITECTURE DETAILS

We use four layers for both the Transformer Encoder and Transformer Decoder. For parameter-efficient adaptation, we insert LoRA modules into the query and value projection layers of the backbone Transformer. The LoRA rank r is set to 16, and the scaling factor α is set to 32. The original weights of the diffusion backbone, VAE, CLIP, and RoBERTa are kept frozen. Only the LoRA parameters and the newly introduced modules are optimized end-to-end.

For TQE, we compute the cosine similarity between the global text feature and each frame-level query feature. We select the top-k text-aligned reference frames, with $k = 3$. Each query attends to its adjacent frames and the selected long-range reference frames. This sparse interaction provides both short-term continuity and text-relevant long-range context without attending to every frame.

A.2 INFERENCE DETAILS

During inference, each referring expression is processed independently. If a video exceeds the maximum input length of the diffusion backbone, we divide it into multiple interleaved subsequences. Each subsequence contains the first frame and samples subsequent frames at a fixed stride with a different temporal offset. Together, these subsequences cover all frames of the original video. We process each subsequence separately and then restore the predictions to their original temporal order. This strategy enables inference on long videos with manageable memory consumption.

For each query, we average its confidence scores across all frames and select the query with the highest video-level score as the final target query. The corresponding masks are resized to the original resolution and thresholded at 0.5 to obtain the final binary mask predictions for each frame.

B MORE ABLATION STUDIES

B.1 ABLATION ON TEMPORAL QUERY EVOLUTION

Table 7 compares sparse and dense temporal attention. Sparse attention improves as k increases from 1 to 3, and the best result is achieved at $k = 3$ with 72.03 $\mathcal{J}\&\mathcal{F}$. This exceeds dense attention by 0.19 points. When k is increased to 4, the performance slightly decreases to 71.94. Using one or two reference frames provides less temporal context and performs worse than dense attention.

Too few references limit context; more references may add irrelevant information. The setting $k = 3$ best balances temporal coverage and selective aggregation. All experiments use $k = 3$.

C TEMPORAL CONSISTENCY ANALYSIS

To evaluate temporal stability on the Ref-YouTube-VOS validation set, we use *temporal fluctuation* to measure the variation in frame-level $\mathcal{J}\&\mathcal{F}$ scores across different temporal gaps. Specifically, let $s_t^{(v)}$ denote the $\mathcal{J}\&\mathcal{F}$ score of frame t in video v. The temporal fluctuation at gap k is defined as

$$\mathrm{TF}(k) = \mathbb{E}_{v,t}\left[\left|s_t^{(v)} - s_{t+k}^{(v)}\right|\right], \tag{11}$$

where $\mathbb{E}_{v,t}$ averages over videos and valid frame pairs. Lower $\mathrm{TF}(k)$ indicates more stable accuracy.

As shown in Fig. 6, the temporal fluctuation of both methods increases with the temporal gap. CGD has lower fluctuation at every evaluated gap, with larger differences at longer gaps.

D QUALITATIVE RESULTS ON CHALLENGING CASES

As shown in Fig. 7, our method produces stable predictions for small targets, dense multi-object interactions, similar distractors, and heavy occlusions. In the first example, although the referred bird

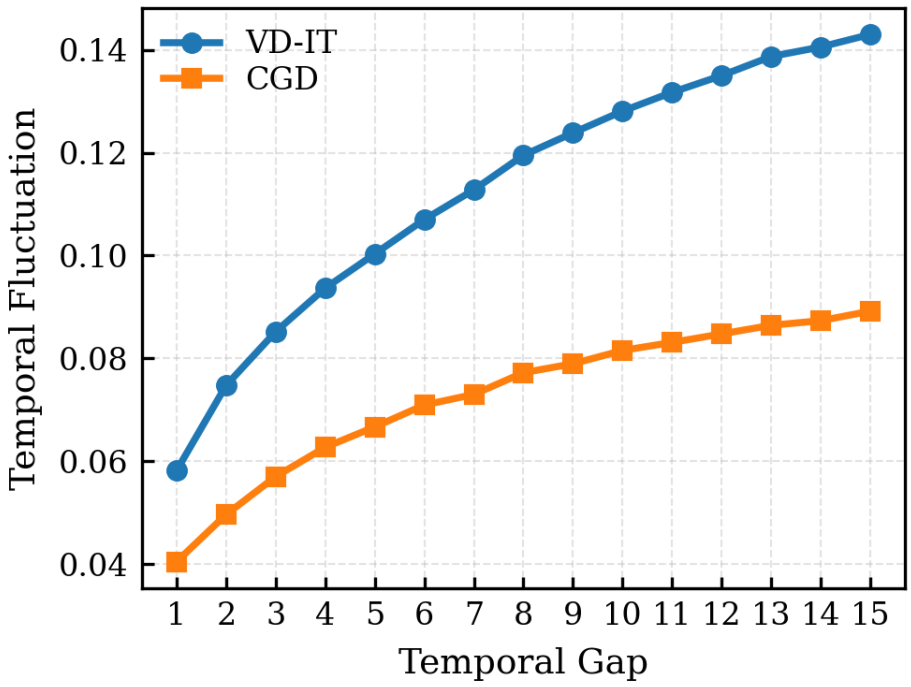

Figure 6: Temporal fluctuation of CGD and VD-IT at different temporal gaps on the Ref-YouTube-VOS. Lower values indicate more stable frame-level segmentation accuracy.



Figure 7: Qualitative comparisons on single object cases, multi object interactions, similar distractors, and heavy occlusions. Our method improves localization and temporal stability.

occupies only a small region and moves over a large background, our method maintains accurate localization without obvious drift. In the second example, multiple zebras with highly similar appearances interact closely, making instance-level discrimination difficult. Our method consistently focuses on the referred target while preserving relatively complete object regions. In the third exam-



Figure 8: Visualization of predicted masks from different variants. DIB mainly improves local details and target discrimination, HSB helps preserve holistic object regions, and CGD combines the advantages of both branches for more complete and accurate mask prediction.

Attention Strategy	J&F	J	F
Sparse ($k = 1$)	71.52	69.36	73.68
Sparse ($k = 2$)	71.79	69.61	73.97
Sparse ($k = 3$)	**72.03**	**69.96**	**74.10**
Sparse ($k = 4$)	71.94	69.80	74.08
Dense	71.84	69.83	73.84

Table 7: Effect of different temporal attention strategies in TQE. Dense attention attends to all frames, whereas sparse attention uses adjacent frames and the top-k text-aligned reference frames.

ple, the referred cat appears together with nearby cats of similar appearance. Our method suppresses responses to these distractors and produces stable masks for the referred instance. In the final example, our masks remain temporally coherent under partial occlusion in a cluttered environment.

E Visual Comparisons of CGD, DIB, and HSB

We compare the masks predicted by CGD, DIB, and HSB in Fig. 8. In both examples, DIB preserves finer details and better suppresses distractors, but sometimes produces incomplete masks. HSB provides broader object coverage but generates coarser boundaries and more responses to nearby regions. CGD combines both branches to improve target discrimination and mask completeness.