Boosting Text-Driven Video Segmentation via Geometry-Aware Distillation

Tianyu Zhu[1]*, Yingping Liang[1]*, Hesong Li[1], and Ying Fu[1]†

Beijing Institute of Technology, Beijing, China
{zhutianyu,liangyingping,lihesong2,fuying}@bit.edu.cn

Abstract. Text-driven Referring Video Object Segmentation (RVOS) aims to locate and segment target objects in videos given natural language. However, existing models are typically trained on 2D image or video datasets with naive segmentation losses, which overlooks the geometric consistency across frames and leads to weak spatial understanding. In this paper, we propose **Geo**metry-enhanced **La**nguage-guided **V**ideo segmentation (**GeoLaV**), a two-stage framework that distills 3D geometric knowledge from images to enhance text-driven video segmentation. In the first stage, we perform monocular geometry pretraining with monocular novel-view synthesis, enabling the model to acquire geometry-consistent visual representations via spatial alignment on large-scale single-image datasets. In the second stage, we introduce geometry-aware distillation and fine-tune the model on video segmentation datasets, transferring 3D structural knowledge from a general 3D prior model. This process reinforces 3D awareness and improves both spatiotemporal coherence and language grounding in segmentation. Extensive experiments show that our method using only image segmentation data already provides notable zero-shot generalization in RVOS. When combined with geometry-aware distillation for fine-tuning on videos, our method achieves state-of-the-art performance across multiple RVOS benchmarks. The code is available at `https://github.com/Tony1882880/GeoLaV`.

Keywords: Referring Video Object Segmentation · Geometry-Aware Learning · 3D Distillation

1 Introduction

Segmenting objects in videos from natural language descriptions is a critical task in multimodal video understanding [4, 56]. Text-driven Referring Video Object Segmentation (RVOS) [10, 14] aligns visual dynamics with linguistic semantics to localize the described target. Traditional video segmentation methods [68, 72], such as mask propagation-based [39] or category-specific approaches [63], rely on predefined labels and lack flexibility in open-world scenarios. Recently, RVOS enables free-form language to specify objects by motion and appearance, supporting diverse applications [49,54]. Advances in vision-language models have further boosted this paradigm, driving progress in multimodal understanding.

* Equal contribution. † Corresponding author.

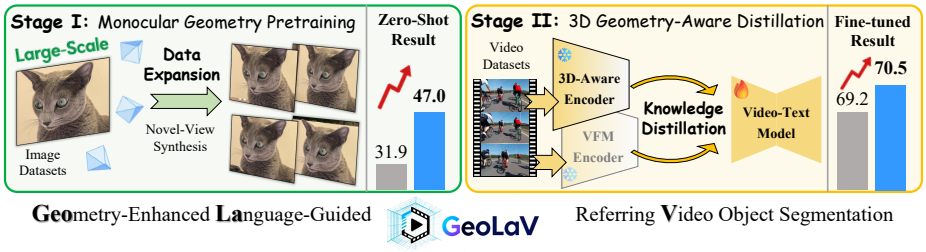

Fig. 1: **Geo**metry-enhanced **La**nguage-guided referring **V**ideo object segmentation (GeoLaV) adopts a two-stage paradigm: **Stage I** performs monocular geometry pretraining via novel-view synthesis to learn geometry-consistent features; **Stage II** applies geometry-aware distillation with 3D teachers for spatiotemporal understanding.

However, the pretraining stage of most existing video segmentation methods [5, 9, 74] remains largely dependent on large-scale 2D image datasets such as COCO [33], due to the scarcity and high cost of video annotations [6, 45]. These image-based pretraining datasets lack cross-frame geometric consistency and viewpoint variations, which are important for learning spatial geometry across views. As a result, models exhibit limited geometric awareness, making it difficult to maintain consistent object representations under viewpoint changes.

A further limitation arises from the fine-tuning paradigm of current video segmentation methods [5, 44], which typically depend on memory banks or temporal attention to implicitly model cross-frame relations. However, they are often trained with naive segmentation losses on datasets of 2D images, overlooking 3D geometric consistency across frames and weakening spatial understanding. Without knowledge of 3D geometry, these models struggle to maintain stable tracking and consistent segmentation under complex camera motion and large object displacement, hindering the accurate modeling of dynamic scenes.

In this paper, we propose a novel two-stage framework, **Geo**metry-enhanced **La**nguage-guided **V**ideo segmentation (**GeoLaV**), which addresses these limitations by explicitly integrating geometric priors into text-driven video segmentation. Specifically, **in the first stage**, we introduce monocular geometry pretraining, where multiple novel-view images are synthesized from a single image under controlled camera motion to form geometry-consistent view sequences in an efficient and scalable manner. Rather than modeling complex temporal dynamics, this stage enforces cross-view structural consistency and encourages viewpoint-robust feature learning from synthetic data. With the aid of a visual foundation model encoder (*e.g.*, DINOv2 [40], DINOv3 [48]), the synthesized views are supervised in a distillation-style manner, encouraging the model to learn geometry-consistent representations across views and providing a geometry-aware initialization for subsequent fine-tuning on real video datasets.

Furthermore, **in the second stage**, we perform 3D geometry-consistent distillation to further enhance the model's understanding of spatial structure and motion. We incorporate general 3D models (*e.g.*, VGGT [52], π^3 [53]) as teachers

to provide 3D geometry priors and transfer geometry-aware representations into the segmentation framework. Through this distillation process, the model learns to maintain the geometric consistency of objects across frames, improving its capability to track targets under complex motions and viewpoint changes. This stage complements the monocular pretraining and enables our GeoLaV to jointly leverage semantic understanding and 3D geometric priors for text-driven video segmentation. Experiments demonstrate the state-of-the-art performance of our method across benchmarks.

In summary, our main contributions are as follows:

- We propose **GeoLaV**, a two-stage framework that integrates single-view novel-view synthesis and 3D geometric priors to learn cross-frame consistency for text-driven video segmentation.
- In **Stage I**, we perform monocular geometry pretraining via novel-view synthesis to learn geometry-consistent and temporally coherent features from large-scale 2D data.
- In **Stage II**, we introduce 3D geometry-aware distillation to transfer 3D structural priors from general 3D models, enhancing spatial and temporal coherence across frames.

2 Related Work

Text-Driven Referring Video Segmentation. Text-driven RVOS [5, 12, 14, 15,36,58] aims to segment target objects in videos given natural language descriptions. Early methods relied on two-stream architectures combining visual and textual features, later evolving into transformer-based frameworks that enable end-to-end multimodal learning [3, 46]. MTTR [3] first unified vision-language reasoning in a single Transformer, achieving strong results on A2D-Sentences [11] and Ref-YouTube-VOS [45]. SOC [36] introduced semantic-assisted object clustering to enhance inter-frame consistency, while MUTR [60] employed a unified temporal transformer for long-range reasoning. To handle motion-sensitive expressions, DsHmp [14] decoupled static and motion perception for finer temporal grounding. Recent lightweight frameworks such as SAMWISE [5] further leverage strong segmentation priors from SAM2 [44] through temporal adapters and memory-guided fusion, achieving better performance under a compact design.

With the development of deep learning [1,7,8,17,24–26,50,51,67,69–71], large vision-language models (Large VLMs) have been explored for text-driven RVOS but remain computationally demanding. LISA [22] employed a large multimodal LLM to reason about segmentation masks from natural prompts, VISA [59] extended such reasoning to video through a mask decoder, and GLUS [32] unified global-local reasoning within a single model. Despite their impressive performance, these Large VLM-based methods rely heavily on large-scale pretraining and contain enormous parameters and computation costs, leading to slow inference. In contrast, non-Large VLM-based methods are more efficient but often overlook geometric cues, limiting spatial consistency across frames.

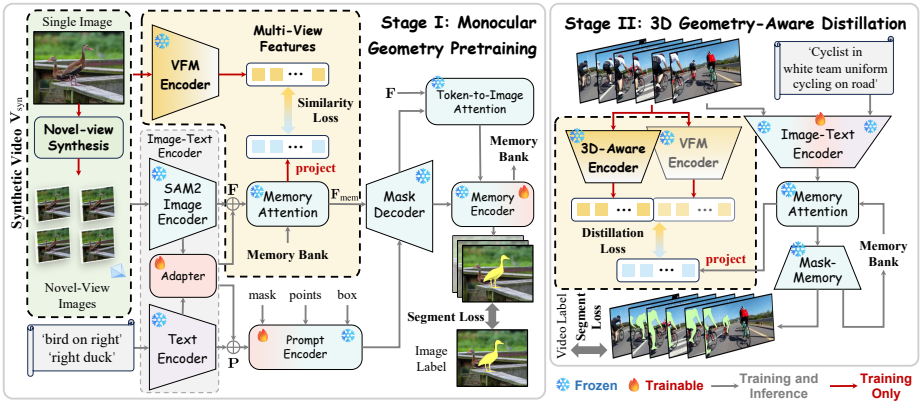

Fig. 2: Overview of our GeoLaV. In the first stage, the monocular geometry pretraining process synthesizes novel views from single images and employs a visual foundation model encoder (*e.g.*, DINOv3 [48]) to learn inter-frame semantic representations. In the second stage, the geometry-aware distillation process utilizes a frozen 3D-aware teacher (*e.g.*, π^3 [53]) to inject 3D geometric priors into the video model, enhancing its spatiotemporal segmentation capability. Stage II shares the same backbone architecture as Stage I (illustrated in a simplified form in the figure) and is initialized with the pretrained weights from Stage I.

3D-Aware Geometry Representation Learning. Recent progress in 3D geometry representation learning [28–30, 55, 73] has advanced from monocular depth estimation to large-scale 3D-aware foundation models. Transformer-based approaches, such as DPT [43], Depth Anything v2 [61, 62], and Marigold [20], demonstrate that large-scale monocular supervision from synthetic and pseudo-labeled images enables encoders to learn robust depth priors from single images. Meanwhile, works such as 3D-to-2D distillation [35] show that geometric knowledge from 3D representations can be effectively transferred to 2D perception tasks. Building on this paradigm, recent 3D foundation models, including MASt3R [23], OpenScene [42], VGGT [52], and π^3 [53], further extend geometric reasoning toward multi-view consistency and open-vocabulary 3D understanding by jointly modeling depth, camera pose, and semantic features.

Moreover, recent findings [19] reveal that multimodal large language models lack intrinsic 3D awareness and benefit significantly from geometry-informed supervision. These insights motivate integrating 3D priors into text-driven video segmentation to enhance spatial coherence and temporal understanding.

3 Method

In this section, we first outline the formulation and motivation of our **GeoLaV**. As shown in Fig. 2, we introduce the Monocular Geometry Pretraining (MGP) stage to capture cross-view geometric consistency from synthesized videos, and

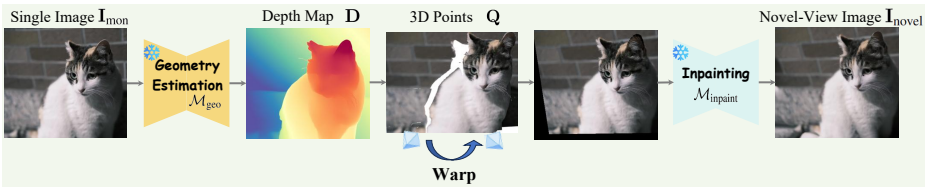

Fig. 3: Pipeline of novel-view image synthesis. Starting from a single RGB input, we estimate depth using a feed-forward geometry model, project pixels into 3D space, and apply virtual camera transformations to obtain new viewpoints. An occlusion mask and an inpainting network are used to refine the rendered views, ensuring geometric coherence and visual completeness.

the Geometry-Aware Distillation (GAD) stage to inject 3D geometric priors into text-driven referring video object segmentation.

3.1 Formulation and Motivation

Given a video sequence $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^{T}$, where each frame $\mathbf{I}_t$ denotes an RGB image, and a referring sentence $\mathbf{W} = \{w_i\}_{i=1}^{L}$ describing the target object, text-driven referring video segmentation aims to predict a sequence of binary masks $\mathbf{M} = \{\mathbf{M}_t\}_{t=1}^{T}$ indicating the object regions in all frames. Formally, the task can be formulated as:

$$\mathbf{M} = \mathcal{G}(\mathbf{V}, \mathbf{W}), \tag{1}$$

where $\mathcal{G}$ denotes the segmentation function integrating visual features from $\mathbf{V}$ and linguistic semantics from $\mathbf{W}$ to produce coherent masks across frames.

However, there are still two main challenges. First, existing RVOS methods are typically pretrained on RefCOCO/+/g [38,66], which consist solely of static images. Such pretraining provides only spatial correspondence between text and individual frames but fails to expose the model to cross-view geometric variation. To address this, we synthesize geometry-consistent multi-view images from a single frame to form pseudo videos, allowing the model to perceive continuous visual transitions. We then introduce a visual foundation model (VFM) encoder to learn cross-frame semantic relationships within these generated sequences, enabling the model to build a more robust spatio-temporal understanding before fine-tuning on real videos.

Second, existing frameworks lack explicit modeling of 3D geometric consistency between frames, which is crucial for accurately tracking moving objects in dynamic scenes. To overcome this, we incorporate a 3D-aware encoder that learns frame-to-frame geometric relationships by aligning with features from pretrained 3D perception models. This higher-dimensional geometric supervision enables the model to capture object structure and motion beyond 2D appearance cues, resulting in more stable and geometry-consistent video segmentation.

3.2 Monocular Geometry Pretraining

Overall Architecture. Given the large-scale availability of image datasets, most pretraining frameworks rely solely on static 2D images, which overlook the temporal relationships crucial for video understanding. To bridge this gap, we introduce a monocular geometry pretraining framework that explicitly learns inter-frame consistency by converting single images into geometry-consistent synthetic videos. Each synthetic video $\mathbf{V}_{\text{syn}} = \{\mathbf{I}^t\}_{t=1}^{T}$ is constructed by combining the original image with its 3D geometry-consistent novel-view renderings, generated through the synthesis strategy described later. These videos serve as training inputs, allowing the model to capture semantics across frames while leveraging large-scale image-based data.

As shown in Figure 2, the video $\mathbf{V}_{\text{syn}}$ is processed by the frozen SAM2 [44] image encoder $\mathcal{E}_{\text{img}}$, while the textual description $\mathbf{W}$ is passed through the frozen text encoder $\mathcal{E}_{\text{txt}}$. Both encoders are hierarchical with progressive downsampling. Let l denote the encoder level and t the frame index; the per-level visual and textual features are denoted as

$$\{\mathbf{F}^{t,l}\}_{l=1}^{L} = \mathcal{E}_{\text{img}}(\mathbf{I}^t), \qquad \{\mathbf{P}^{l}\}_{l=1}^{L} = \mathcal{E}_{\text{txt}}(\mathbf{W}). \tag{2}$$

A trainable cross-modal adapter $\mathcal{A}_\theta$ is inserted at each level to align the two modalities in a shared representation space, while keeping both encoders frozen:

$$\mathbf{F}'^{\,t,l} = \mathcal{A}_\theta^l\big(\mathbf{F}^{t,l},\, \mathbf{P}^{l}\big), \quad l = 1, \ldots, L, \quad t = 1, \ldots, T. \tag{3}$$

The aligned features are passed through the frozen memory attention module to obtain memory-enhanced representations $\mathbf{F}_{\text{mem}}^{t,l}$. The prompt encoder combines frozen SAM2 [44] weights with a small set of trainable parameters to encode task prompts. The outputs of the prompt encoder and memory attention are fed into the mask decoder, whose predictions are refined by the memory encoder to produce segmentation masks. Meanwhile, the memory encoder continuously updates a memory bank, which stores temporal information and is reused by the memory attention module in subsequent frames.

In parallel, the synthesized video is also processed by a visual foundation model (VFM) encoder, such as DINOv3 [48], to extract features that provide semantic supervision. Since the alignment is applied before prompt-conditioned decoding, it operates on global frame-level representations rather than object-specific features. The overall training objective is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{sim}}, \tag{4}$$

where segmentation loss $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{mask}}$ combines Dice and mask losses following previous works, and similarity loss $\mathcal{L}_{\text{sim}}$ aligns the VFM features with the projected memory-attention features (detailed later). This design enables the model to jointly learn spatial semantics and temporal consistency from large-scale image data without requiring real video annotations.

Novel-View Images Synthesis. To preserve the 3D geometric fidelity of the original image during synthetic view-sequence generation, we design a geometry-consistent multi-view synthesis framework, as illustrated in Fig. 3. Inspired by

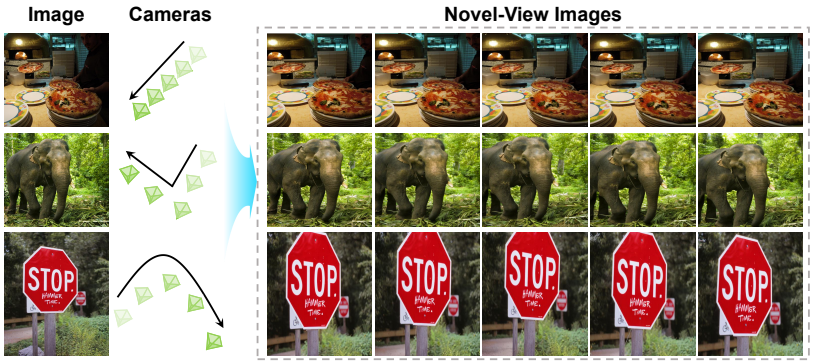

Fig. 4: Examples of synthesized novel-view sequences for our Monocular Geometry Pretraining (MGP). Each row corresponds to a continuous camera trajectory (*e.g.*, linear, piecewise-linear, and curved motion). Adjacent frames are generated through physically valid rigid transformations in 3D space, preserving consistent geometric structure and depth–camera correspondence across views.

L2M [31], we employ recent feed-forward 3D geometry estimation models (*e.g.*, π^3 [53], VGGT [52]) that directly infer scene geometry from a monocular RGB image. These models predict scale-invariant depth or point maps in a reference-free manner, providing a reliable geometric foundation for synthesizing novel views.

Given a monocular image $\mathbf{I}_{\text{mon}}$, the network estimates a dense depth map $\mathbf{D}$. To simulate the inherent scale ambiguity in monocular reconstruction, we introduce a random scaling factor a and a bias term b, *i.e.*,

$$\mathbf{D} = a\mathcal{M}_{\text{geo}}(\mathbf{I}_{\text{mon}}) + b, \tag{5}$$

where $\mathcal{M}_{\text{geo}}$ denotes the chosen 3D geometry estimation model. Although the predicted depth is not metrically precise, it preserves consistent 3D geometry with the original monocular image, ensuring that synthesized frames remain geometrically aligned with the source view.

Using the predicted depth, the image is lifted into 3D space by projecting each pixel (u, v) into 3D coordinates through a sampled intrinsic matrix $\mathbf{K}$, forming a dense and geometry-consistent point set $\mathbf{Q} = \{(X, Y, Z)\}$. As shown in Fig. 4, instead of sampling independent camera poses, we generate a short continuous camera trajectory across frames. The trajectory may follow linear, piecewise-linear, or curved motion patterns to increase diversity while preserving geometric consistency. Starting from an initial pose, subsequent poses are obtained via small incremental rigid transformations under a physically valid camera projection model, forming a coherent view sequence with real geometric correspondence between adjacent frames.

The transformed 3D points are then reprojected into the image plane to produce geometry-consistent novel-view renderings $\mathbf{I}_{\text{novel}}$. To mitigate occlusion and disocclusion artifacts, an occlusion mask $\mathbf{M}$ is computed and an inpainting

network $\mathcal{M}_{\text{inpaint}}$ is adopted to fill missing regions:

$$\mathbf{I}_{\text{syn}} = \mathcal{M}_{\text{inpaint}}(\mathbf{I}_{\text{novel}}, \mathbf{M}), \tag{6}$$

where $\mathcal{M}_{\text{inpaint}}$ restores invisible areas based on surrounding contextual cues. The generated view sequences maintain consistent structure, illumination, and depth–camera correspondence across frames, establishing a geometry-coherent foundation for large-scale pretraining.

Our geometry-based pipeline is deterministic and computationally efficient. A short multi-frame sequence can be synthesized within seconds, enabling scalable data expansion for geometry-aware pretraining on large image datasets.

Feature Projection Module. As shown in the left part of Fig. 2, the features extracted from the SAM2 [44] memory attention branch and the VFM encoder (*e.g.*, DINOv3 [48]) differ in both dimensionality and distribution. To achieve consistent representation learning, we introduce a lightweight projection head that transforms the memory-attention features into the same embedding space as the VFM representations. This projection head is implemented as a two-layer MLP with GELU activation,

$$\mathbf{F}_{\text{proj}} = \mathcal{P}_\theta(\mathbf{F}_{\text{mem}}) = \mathbf{W}_2\, \sigma(\mathbf{W}_1 \mathbf{F}_{\text{mem}}), \tag{7}$$

where $\mathcal{P}_\theta$ denotes the projection function parameterized by learnable weights $\mathbf{W}_1$ and $\mathbf{W}_2$, and $\sigma(\cdot)$ represents the GELU activation. The output dimension of this projection is dynamically matched to the feature dimension of the selected VFM encoder, ensuring compatibility across different visual foundation backbones. This projection is used only during training and introduces no additional computation or parameters at inference.

To inject structural and geometry-aware priors from the 3D and VFM encoders into the SAM2 [44] backbone, we perform feature alignment at the memory-attention level during fine-tuning. Since this representation is computed before prompt-conditioned mask decoding, it tends to capture general feature rather than object-specific features tied to a particular textual query. We employ a cosine similarity loss between the projected feature $\mathbf{F}_{\text{proj}}$ and the corresponding VFM feature $\mathbf{F}_{\text{VFM}}$, *i.e.*,

$$\mathcal{L}_{\text{sim}} = 1 - \text{Sim}(\mathbf{F}_{\text{proj}}, \mathbf{F}_{\text{VFM}}) = 1 - \frac{\mathbf{F}_{\text{proj}} \cdot \mathbf{F}_{\text{VFM}}}{\|\mathbf{F}_{\text{proj}}\|\, \|\mathbf{F}_{\text{VFM}}\|}, \tag{8}$$

where $\mathcal{L}_{\text{sim}}$ denotes the cosine similarity loss, $\text{Sim}(\cdot)$ computes cosine similarity, and $\|\cdot\|$ represents the L_2 norm. The alignment is therefore applied to backbone memory features prior to prompt-conditioned decoding, while the primary supervision remains the text-driven segmentation objective. The similarity loss serves as an auxiliary regularizer that injects geometry-aware structural cues while compensating for the lack of supervision in synthesized frames.

3.3 Geometry-Aware Distillation

Building upon the pretraining stage, this step focuses on enhancing 3D geometric understanding and temporal stability through geometry-aware distillation. As

illustrated in the right part of Fig. 2, the model is fine-tuned on real video datasets so that it can perceive structural variations across frames and maintain consistent object representations.

Each video sequence $\mathbf{V}_{\text{real}} = \{\mathbf{I}^t\}_{t=1}^{T}$ is paired with text descriptions and processed by a 3D-aware encoder, such as π^3 [53], together with a VFM encoder. The 3D-aware encoder provides higher-dimensional geometric priors that capture structural and depth relationships across frames, while the VFM encoder contributes semantic priors that describe object appearance and category. By combining complementary cues, the model learns to perceive objects with stronger spatial coherence and achieves more stable tracking in dynamic scenes.

Similar to the monocular geometry pretraining stage, two projection heads are introduced to align the feature spaces of the encoders with the memory representation. Each projection head is implemented as a two-layer MLP with GELU activation,

$$\mathbf{F}_{\text{proj}}^{(k)} = \mathcal{P}_{\theta}^{(k)}\big(\mathbf{F}_{\text{mem}}\big) = \mathbf{W}_2^{(k)}\,\sigma\big(\mathbf{W}_1^{(k)}\mathbf{F}_{\text{mem}}\big), \quad k \in \{\text{3D, VFM}\}. \tag{9}$$

All notations are consistent with those defined in Sec. 3.2. Their output dimensions are matched to the corresponding encoders, and both heads are used only during training without affecting inference.

The supervision is established by cosine similarity losses computed between the projected memory features and the outputs of each encoder, with the overall distillation objective defined as

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{VFM}} = \sum_{k \in \{\text{3D, VFM}\}} \big[1 - \text{Sim}\big(\mathbf{F}_{\text{proj}}^{(k)}, \mathbf{F}_k\big)\big], \tag{10}$$

where the similarity function $\text{Sim}(\cdot)$ follows the definition in Sec. 3.2. Through this dual-branch distillation, the backbone representation is regularized to encode richer geometric structure across frames. The text-conditioned mask prediction remains guided by the segmentation objective, ensuring that refer-object specificity is preserved while benefiting from geometry-aware cues.

4 Experiments

In this section, we evaluate our GeoLaV through comprehensive experiments. We first outline the experimental setup, followed by the main results on multiple referring video segmentation benchmarks. Finally, we discuss the effects of each design and the benefits introduced by geometry pretraining. Additional experiments and visualizations are provided in the supplementary material.

4.1 Experiment Setup

Datasets. Following the previous works [5, 14], we evaluate our approach on three referring video segmentation benchmarks, *i.e.*, Ref-Youtube-VOS [45], Ref-DAVIS17 [21], and MeViS [6]. Ref-Youtube-VOS extends the YouTube-VOS

Table 1: Quantitative Comparison of our Non-Large VLM-based method with state-of-the-art RVOS approaches on Ref-Youtube-VOS [45], Ref-DAVIS17 [21], and MeViS [6]. For non-large VLM-based methods, the best results are shown in **Bold** and the second-best results are Underlined. We also include large VLM-based methods for reference, where the best results are shown in **Bold**. † indicates methods evaluated with SAM2 [44] integration via GroundingDINO [34].

Method	Total Params	Ref-YouTube-VOS			Ref-DAVIS17			MeViS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
***Large VLM**-based*										
LISA [22]	7B	53.9	53.4	54.3	64.8	62.2	67.3	37.2	35.1	39.4
VISA-7B [59]	7B	61.5	59.8	63.2	**69.4**	**66.3**	**72.5**	43.5	40.7	46.3
VISA-13B [59]	13B	63.0	61.4	64.7	67.0	73.8	70.4	44.5	41.8	47.1
One-Token-Seg-All	3.8B	61.7	60.2	63.3	67.7	63.8	71.5	42.3	39.4	45.2
VideoGLaMM	3.8B	-	-	-	-	-	-	45.2	42.1	48.2
GLUS^S	7B	66.6	65.0	68.3	-	-	-	50.3	47.5	53.2
GLUS^A [32]	7B	**67.3**	**65.5**	**69.0**	-	-	-	**51.3**	**48.5**	**54.2**
***Non-Large VLM**-based*										
MTTR [3]	-	55.3	54.0	56.6	-	-	-	30.0	28.8	31.2
TCE-RVOS [15]	-	59.6	58.3	60.8	59.4	56.5	62.4	-	-	-
ReferFormer [58]	237M	62.9	61.3	64.6	61.1	58.1	64.1	31.0	29.8	32.2
SOC [36]	220M	66.0	64.1	67.9	64.2	61.0	67.4	-	-	-
OnlineRefer [57]	232M	63.5	61.6	65.5	64.8	61.6	67.7	32.3	31.5	33.1
LMPM [6]	195M	-	-	-	-	-	-	37.2	34.2	40.2
DsHmp [14]	339M	67.1	65.0	69.1	64.9	61.7	68.1	-	-	-
DsHmp [14]	272M	-	-	-	-	-	-	46.4	43.0	49.8
MUTR [60]	250M	67.5	65.4	69.6	66.4	62.8	70.0	-	-	-
GroundingDINO [34] †	240M	57.5	55.6	59.5	66.4	62.8	69.9	37.7	34.9	40.5
LAVT [65]	239M	65.8	63.6	67.9	-	-	-	-	-	-
SSA [41]	474M	64.3	62.2	66.4	67.3	64.0	70.7	48.6	44.0	53.2
SAMWISE [5]	150M	67.2	65.2	69.3	68.5	65.6	71.5	48.3	45.4	51.2
SAMWISE [5]	202M	69.2	67.8	70.6	70.6	67.4	73.5	49.5	46.6	52.4
ReferDINO [27]	230M	69.3	67.0	71.5	68.9	65.1	72.9	49.3	44.7	**53.9**
GeoLaV (Ours)	202M	**70.5**	**69.1**	**71.8**	**72.5**	**69.9**	**75.2**	**50.0**	**47.4**	52.9

benchmark by adding textual descriptions, containing 3,978 high-resolution videos and about 15K language expressions that refer to target objects across time. Ref-DAVIS17 builds upon the DAVIS17 dataset, enriching 90 videos with over 1.5K linguistic annotations, supporting referring-expression segmentation in scenes with multiple moving objects and complex backgrounds. MeViS includes 2,006 videos and approximately 28K motion-centric expressions that capture complex motion patterns and emphasize motion description rather than static cues.

Implementation Details. Our model architecture is basically built upon the SAMWISE framework [5], with modifications to incorporate geometry-aware feature learning. We employ DINOv3-ViT-L [48] as VFM encoder and π^3 [53] as 3D-aware encoder. **In the first stage,** we synthesize five novel-view images from each COCO [33] single-view input to form pseudo videos. Following previous works [5, 13, 36], the pretraining is performed on RefCOCO/+/g [38, 66] for 6 epochs, with an initial learning rate of 1×10^{-4} decayed to 2×10^{-5}. **In the second stage,** we fine-tune on video data from Ref-Youtube-VOS [45] for 4 epochs with a learning rate of 2×10^{-6}. The trained model is evaluated on Ref-Youtube-VOS and Ref-DAVIS17 [21], and for MeViS [6] we train for 2 epochs

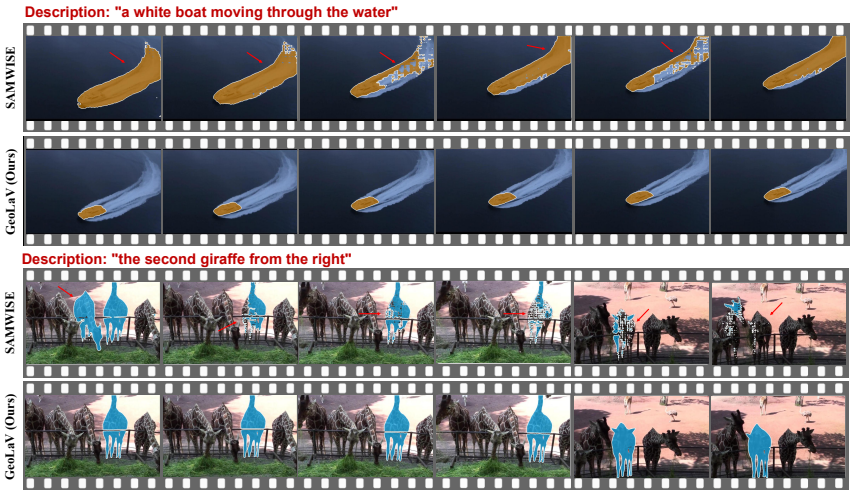

Fig. 5: Qualitative comparison with SAMWISE [5]. It represents the state-of-the-art method and serves as the base architecture of our model. Our GeoLaV more accurately identifies target objects and maintains stable tracking across challenging video frames.

under the same protocol. All experiments use the Adam optimizer on 4 NVIDIA A100 GPUs (80GB GPU memory each).

Evaluation Metrics. We adopt evaluation metrics, including region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$), and their mean score ($\mathcal{J}$ & $\mathcal{F}$). For MeViS [6] and Ref-Youtube-VOS [45], evaluations are performed via the official challenge servers, while Ref-DAVIS17 [21] is assessed using the official evaluation code.

4.2 Main Results

We present the main results to comprehensively evaluate GeoLaV. We first introduce the compared methods from both non-large and large VLM-based RVOS approaches. Then, we compare the overall performance on three benchmarks, showing consistent gains in segmentation quality.

Comparison Methods. We compare GeoLaV with a wide range of RVOS approaches. The non-large VLM-based group includes MTTR [3], TCE-RVOS [15], ReferFormer [58], SOC [36], OnlineRefer [57], LMPM [6], DsHmp [14], MUTR [60], GroundingDINO [34], SSA [41], SAMWISE [5], and ReferDINO [27], all having comparable parameter scales to our method. In addition, we include large VLM-based approaches such as LISA [22], VISA-7B/13B [59], One-Token-Seg-All [2], VideoGLaMM [37], and GLUS$^{S/A}$ [32] as extra references for comparison on the same benchmarks. All reported results (except GroundingDINO) of these methods are directly taken from their original papers for fair comparison. For GroundingDINO, we report its performance when combined with SAM2 [44], as it is originally designed for object detection.

Quantitative Comparison of Full Framework. As shown in Tab. 1, GeoLaV achieves state-of-the-art performance across all three benchmarks. Among

non-large VLM-based methods, our approach consistently ranks first on Ref-Youtube-VOS [45] (70.5 $\mathcal{J}\&\mathcal{F}$), Ref-DAVIS17 [21] (72.5 $\mathcal{J}\&\mathcal{F}$), and MeViS [6] (50.0 $\mathcal{J}\&\mathcal{F}$), surpassing the previous best results by 1.2, 1.9, and 0.5 points, respectively. Compared with recent strong competitors such as SAMWISE [5] and ReferDINO [27], GeoLaV demonstrates notable improvements in both region similarity and contour accuracy, highlighting the benefits of incorporating geometric priors and temporally consistent reasoning. Furthermore, despite using a much smaller model size than large VLM-based approaches (*e.g.*, GLUS^A [32] with 7B parameters), GeoLaV achieves highly competitive results, demonstrating the efficiency and effectiveness of our framework.

Qualitative Results of Full Framework. Since SAMWISE [5] represents the state-of-the-art and serves as the base architecture of our model, we conduct a direct visual comparison between the two. Additional visual comparisons are provided in the supplementary materials. Our GeoLaV produces more accurate and temporally stable segmentation across diverse scenes. In the first example of Fig. 5, SAMWISE mistakenly segments the white wake behind the boat as part of the target, while our method precisely captures only the boat region. In the second example, SAMWISE fails to consistently track the rightmost giraffe across frames, whereas our approach maintains stable and coherent masks over time. These results demonstrate that incorporating geometric priors effectively enhances spatial precision and temporal consistency, enabling fine-grained segmentation across complex video sequences.

4.3 Discussions

We analyze our GeoLaV from four aspects. We first evaluate the generalization of the geometry-consistent data expansion via zero-shot evaluation after pretraining, then examine the contributions of MGP and GAD. Finally, we analyze geometric accuracy and visualize learned features to understand distillatiWon.

Zero-Shot Evaluation after Pretraining. We perform zero-shot evaluation to validate the effectiveness of our geometry-consistent data expansion strategy in extending image pretraining to video segmentation. The model is trained only on RefCOCO/+/g [38, 66] and directly tested on Ref-Youtube-VOS [45] and MeViS [6] without any video fine-tuning. For comparison, we use SAMWISE [5], which represents the current state-of-the-art method and serves as the base architecture of our model. We further include other video-based methods that provide publicly available image-pretrained weights, such as LMPM [6], DsHmp [14], and LAVT [65], as well as referring image segmentation methods including DMMI [16], ReMamber [64], and DETRIS [18]. As shown in Tab. 2, GeoLaV achieves 47.0 on Ref-Youtube-VOS and 31.6 on MeViS, improving upon SAMWISE by $+$**15.1** and $+$**5.2** $\mathcal{J}\&\mathcal{F}$, indicating notable generalization from image pretraining to unseen video domains.

Effectiveness of the Monocular Geometry Pretraining. As shown in Tab. 3, introducing Monocular Geometry Pretraining (Stage I) boosts the $\mathcal{J}\&\mathcal{F}$ score from 65.1 to 67.2. Compared with planar homography-based multi-frame

Table 2: Zero-shot performance when trained on images and evaluated on video benchmarks. All models are trained on RefCOCO/+/g [38, 66] only.

Method	Ref-YouTube-VOS [45]			MeViS [6]		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
***Image**-based Model*						
DMMI [16] [ICCV'23]	9.6	6.6	7.5	7.1	6.9	7.3
ReMamber [64] [ECCV'24]	35.1	35.3	35.0	26.5	24.6	28.4
DETRIS [18] [AAAI'25]	17.6	17.9	17.4	16.3	16.1	16.5
***Video**-based Model* (Trained on Images Only)						
LMPM [6] [ICCV'23]	7.1	7.0	7.1	11.8	11.6	12.0
DsHmp [14] [CVPR'24]	7.1	7.0	7.1	11.8	11.6	12.0
LAVT [65] [TPAMI'25]	28.6	28.9	28.4	21.8	22.2	21.6
SAMWISE [5] [CVPR'25]	31.9	31.2	32.6	26.4	24.6	28.1
GeoLaV (Ours)	**47.0**	**46.9**	**47.2**	**31.6**	**29.6**	**33.6**

Table 3: Ablation study on the effectiveness of Monocular Geometry Pretraining (MGP) and Geometry-Aware Distillation (GAD). All models are first pretrained on image datasets and then fine-tuned on video benchmarks. Planar augmentation denotes homography-based 2D transformations (*e.g.*, translation, scaling, rotation, cropping) that generate pseudo multi-frame inputs without geometric modeling.

Setting	Stage I	Stage II	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Vanilla Model	Base	Base	65.1	63.1	66.9
+ Planar Augmentation	Planar	Base	66.5	64.5	67.6
+ MGP (Ours)	MGP	Base	67.2	66.1	68.0
+ GAD (Ours)	Base	GAD	67.5	66.2	68.9
Full Model (Ours)	MGP	GAD	**70.5**	**69.1**	**71.8**

augmentation, which yields only moderate gains, our depth-based warping strategy brings consistently larger improvements. This indicates that the performance gain does not merely stem from increased frame exposure, but from enforcing cross-view geometric consistency during pretraining. By synthesizing geometry-consistent novel views from single images, the encoder learns structure-aware representations that are more robust to viewpoint variations, providing a stronger initialization for subsequent video fine-tuning.

Effectiveness of the Geometry-Aware Distillation. Tab. 3 shows that applying Geometry-Aware Distillation (Stage II) improves $\mathcal{J}\&\mathcal{F}$ from 65.1 to 67.5. By aligning the video model with 3D-aware teacher features during fine-tuning, this stage injects structural and depth-aware priors that promote geometry-consistent representations across frames. Compared with generic feature supervision, such priors help maintain more stable object representations under viewpoint changes and motion. Moreover, combining Stage II with Stage I leads to substantially larger gains, suggesting that geometry-consistent pretraining provides a better initialization for absorbing 3D structural cues.

Geometric Accuracy Analysis. To validate that our synthesis preserves true 3D geometry, we compare it with homography-based planar transformations in Fig. 6. Planar augmentation applies 2D warping without modeling scene depth, often causing distorted shapes and incorrect occlusion under viewpoint changes. In contrast, our method reconstructs a 3D representation and reprojects it under

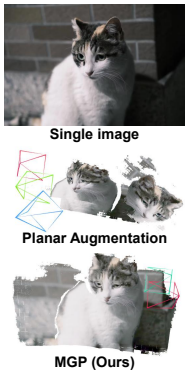

Fig. 6: Geometric consistency comparison: planar vs. ours.

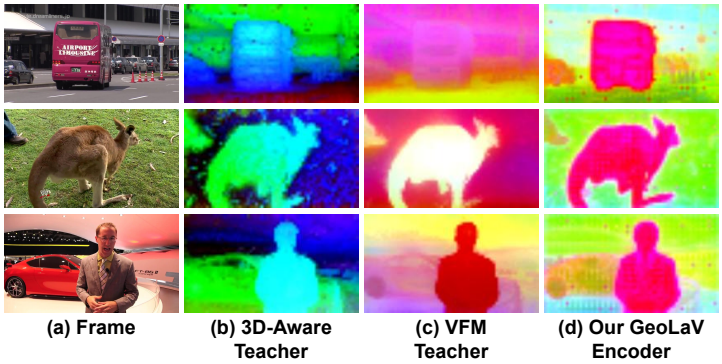

Fig. 7: Visualization of feature representations after geometry-aware distillation via PCA [47]. GeoLaV integrates geometric and semantic cues from dual teachers, yielding more distinguishable object-level features.

continuous camera motion, preserving object geometry and depth ordering across views. This demonstrates that our synthesis maintains physically consistent 3D structure rather than performing purely 2D augmentation.

Feature Visualization. Beyond geometric consistency at the image level, we further analyze how the learned representations encode structural and semantic information. To understand what GeoLaV learns from the two teachers, we visualize the encoder features extracted from the 3D-aware and VFM encoders in Fig. 7. The 3D-aware teacher provides depth and geometric cues, while the VFM teacher captures clear semantic boundaries and object-level understanding. After geometry-aware distillation, our encoder integrates both advantages, preserving geometric structure and maintaining sharp semantic edges. This shows that the dual-teacher design produces geometry-preserving and semantically consistent representations for precise video segmentation.

5 Conclusion

In this paper, we introduce **GeoLaV**, a two-stage framework that enhances text-driven video object segmentation through geometry-guided learning. Our GeoLaV addresses the limitations of prior 2D-based methods, which lack geometric consistency and spatial awareness. GeoLaV first employs monocular geometry pretraining that synthesizes novel views from single images, enabling geometry-consistent visual representation learning. Then, a geometry-aware distillation stage transfers 3D structural knowledge from general 3D priors, reinforcing spatiotemporal coherence and improving visual-language alignment. Extensive experiments on multiple benchmarks show that GeoLaV achieves state-of-the-art performance, validating the benefit of incorporating geometric reasoning into multimodal video understanding. Future work includes extending this framework toward open-vocabulary and large-scale 3D-aware multimodal segmentation.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62331006), the Fundamental Research Funds for the Central Universities, and the National Natural Science Foundation of China under Grant (625B2026).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. In: NeurIPS. pp. 6833–6859 (2024)
3. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: CVPR. pp. 4985–4995 (2022)
4. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Lin, B., Tang, Z., Yuan, L., Qiao, Y., Lin, D., Zhao, F., Wang, J.: Sharegpt4video: Improving video understanding and generation with better captions. In: NeurIPS. pp. 19472–19495 (2024)
5. Cuttano, C., Trivigno, G., Rosi, G., Masone, C., Averta, G.: Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In: CVPR. pp. 3395–3405 (2025)
6. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: ICCV. pp. 2694–2703 (2023)
7. Dong, W., Wang, Y.C., Yang, C., Sun, C., Li, H., Hua, Z., Wu, Z., Chang, X., Bao, L., Qu, S., et al.: Deep learning enhanced in situ atomic imaging of ion migration at crystalline–amorphous interfaces. Nano Letters **24**(45), 14445–14452 (2024)
8. Fan, J., Zhang, K., Zhao, Y., Liu, Q.: Unsupervised video object segmentation via weak user interaction and temporal modulation. Chinese Journal of Electronics **32**(3), 507–518 (2023)
9. Fang, H., Wu, P., Li, Y., Zhang, X., Lu, X.: Unified embedding alignment for open-vocabulary video instance segmentation. In: ECCV. pp. 225–241 (2024)
10. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.M.: Actor and action video segmentation from a sentence. In: CVPR (2018)
11. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.M.: Actor and action video segmentation from a sentence. In: CVPR. pp. 5958–5966 (2018)
12. Gu, C., Chen, L., Gu, L., Fu, Y.: Fourier angle alignment for oriented object detection in remote sensing. In: CVPR. pp. 42225–42235 (2026)
13. Han, M., Wang, Y., Li, Z., Yao, L., Chang, X., Qiao, Y.: Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In: ICCV. pp. 13414–13423 (2023)
14. He, S., Ding, H.: Decoupling static and hierarchical motion perception for referring video segmentation. In: CVPR. pp. 13332–13341 (2024)
15. Hu, X., Hampiholi, B., Neumann, H., Lang, J.: Temporal context enhanced referring video object segmentation. In: WACV. pp. 5574–5583 (2024)
16. Hu, Y., Wang, Q., Shao, W., Xie, E., Li, Z., Han, J., Luo, P.: Beyond one-to-one: Rethinking the referring image segmentation. In: ICCV. pp. 4067–4077 (2023)

17. Hua, Z., Qu, S., Yan, L., Dong, W., Zhou, Y., Li, H., Chang, X., Bao, L., Wang, Y., Ying, F., et al.: Deep-learning aided atomic-scale observation of anisotropic melting of the charge density wave in tas2. Small **21**(45), e07496 (2025)
18. Huang, J., Xu, Z., Liu, T., Liu, Y., Han, H., Yuan, K., Li, X.: Densely connected parameter-efficient tuning for referring image segmentation. In: AAAI. pp. 3653–3661 (2025)
19. Huang, X., Wu, J., Xie, Q., Han, K.: Mllms need 3d-aware representation supervision for scene understanding. arXiv preprint arXiv:2506.01946 (2025)
20. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR. pp. 9492–9502 (2024)
21. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: ACCV. pp. 123–141 (2018)
22. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
23. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV. pp. 71–91 (2024)
24. Li, H., Fu, Y.: Fcdfusion: A fast, low color deviation method for fusing visible and infrared image pairs. Computational Visual Media **11**(1), 195–211 (2025)
25. Li, H., Wu, Z., Shao, R., Fu, Y.: Statistical characteristic-guided denoising for rapid high-resolution transmission electron microscopy imaging. In: CVPR. pp. 34050–34060 (2026)
26. Li, H., Wu, Z., Shao, R., Zhang, T., Fu, Y.: Noise calibration and spatial-frequency interactive network for stem image enhancement. In: CVPR. pp. 21287–21296 (June 2025)
27. Liang, T., Lin, K.Y., Tan, C., Zhang, J., Zheng, W.S., Hu, J.F.: Referdino: Referring video object segmentation with visual grounding foundations. In: ICCV. pp. 20009–20019 (2025)
28. Liang, Y., Fu, Y.: Relation-guided adversarial learning for data-free knowledge transfer. IJCV **133**(5), 2868–2885 (2025)
29. Liang, Y., Fu, Y., Hu, Y., Shao, W., Liu, J., Zhang, D.: Flow-anything: Learning real-world optical flow estimation from large-scale single-view images. IEEE TPAMI **47**(10), 8435–8452 (2025)
30. Liang, Y., Fu, Y., Liu, J., Zhang, D.: Lift3dreamer: Boosting text-driven novel view synthesis via lifted 3d inpainting model from single images. Fundamental Research (2026)
31. Liang, Y., Hu, Y., Shao, W., Fu, Y.: Learning dense feature matching via lifting single 2d image to 3d space. In: ICCV. pp. 6621–6631 (2025)
32. Lin, L., Yu, X., Pang, Z., Wang, Y.X.: Glus: Global-local reasoning unified into a single large language model for video segmentation. In: CVPR. pp. 8658–8667 (2025)
33. 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: ECCV. pp. 740–755 (2014)
34. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55 (2024)
35. Liu, Z., Qi, X., Fu, C.W.: 3d-to-2d distillation for indoor scene parsing. In: CVPR. pp. 4464–4474 (2021)

36. Luo, Z., Xiao, Y., Liu, Y., Li, S., Wang, Y., Tang, Y., Li, X., Yang, Y.: Soc: Semantic-assisted object cluster for referring video object segmentation. In: NeurIPS. pp. 26425–26437 (2023)
37. Munasinghe, S., Gani, H., Zhu, W., Cao, J., Xing, E., Khan, F.S., Khan, S.: Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In: CVPR. pp. 19036–19046 (2025)
38. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: ECCV. pp. 792–807 (2016)
39. Oh, S.W., Lee, J.Y., Sunkavalli, K., Kim, S.J.: Fast video object segmentation by reference-guided mask propagation. In: CVPR. pp. 7376–7385 (2018)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
41. Pan, F., Fang, H., Li, F., Xu, Y., Li, Y., Benini, L., Lu, X.: Semantic and sequential alignment for referring video object segmentation. In: CVPR. pp. 19067–19076 (2025)
42. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: CVPR. pp. 815–824 (2023)
43. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
44. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
45. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223 (2020)
46. Seo, S.W., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223 (2020)
47. Shlens, J.: A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100 (2014)
48. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
49. Thawakar, O., Naseer, M., Anwer, R.M., Khan, S., Felsberg, M., Shah, M., Khan, F.S.: Composed video retrieval via enriched context and discriminative embeddings. In: CVPR. pp. 26896–26906 (2024)
50. Tian, Y., Fu, Y., Zhang, J.: Transformer-based under-sampled single-pixel imaging. Chinese Journal of Electronics **32**(5), 1151–1159 (2023)
51. Wang, J., Li, H., Wang, X., Fu, Y.: 3d-b2u: Self-supervised fluorescent image sequences denoising. In: CAAI International Conference on Artificial Intelligence. pp. 130–142 (2024)
52. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025)
53. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Scalable permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
54. Wang, Y., Wang, L., Ma, Z., Hu, Q., Xu, K., Guo, Y.: Videodirector: Precise video editing via text-to-video models. In: CVPR. pp. 2589–2598 (2025)
55. Wang, Y., Liang, Y., Hu, Y., Fu, Y.: Robustereo: Robust zero-shot stereo matching under adverse weather. In: ICCV. pp. 25134–25144 (2025)

56. Wang, Y., Song, Y., Xie, C., Liu, Y., Zheng, Z.: Videollamb: Long streaming video understanding with recurrent memory bridges. In: ICCV. pp. 24170–24181 (2025)
57. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: ICCV. pp. 2761–2770 (2023)
58. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022)
59. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: ECCV. pp. 98–115 (2024)
60. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., Dong, H., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation. In: AAAI. pp. 6449–6457 (2024)
61. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR. pp. 10371–10381 (2024)
62. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. NeurIPS pp. 21875–21911 (2024)
63. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV. pp. 5188–5197 (2019)
64. Yang, Y., Ma, C., Yao, J., Zhong, Z., Zhang, Y., Wang, Y.: Remamber: Referring image segmentation with mamba twister. In: ECCV. pp. 108–126 (2024)
65. Yang, Z., Wang, J., Ye, X., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Language-aware vision transformer for referring segmentation. IEEE TPAMI **47**(7), 5238–5255 (2025)
66. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85 (2016)
67. Zhang, T., Fu, Y., Zhang, J., Yan, C.: Deep guided attention network for joint denoising and demosaicing in real image. Chinese Journal of Electronics **33**(1), 303–312 (2024)
68. Zhang, T., Tian, X., Zhou, Y., Ji, S., Wang, X., Tao, X., Zhang, Y., Wan, P., Wang, Z., Wu, Y.: Dvis++: Improved decoupled framework for universal video segmentation. IEEE TPAMI **47**(7), 5918–5929 (2025)
69. Zhang, Y., Lai, Z., Zhang, T., Fu, Y., Zhou, C.: Unaligned rgb guided hyperspectral image super-resolution with spatial-spectral concordance. IJCV **133**(9), 6590–6610 (2025)
70. Zhang, Y., Zhang, T., Nie, J., Fu, Y.: Enhancing unregistered hyperspectral image super-resolution via unmixing-based abundance fusion learning. In: CVPR. pp. 41573–41583 (2026)
71. Zhang, Y., Zhang, T., Nie, J., Fu, Y.: Real noise decoupling for hyperspectral image denoising. In: AAAI. pp. 12925–12933 (2026)
72. Zhou, Y., Zhang, T., Ji, S., Yan, S., Li, X.: Improving video segmentation via dynamic anchor queries. In: ECCV. pp. 446–463 (2024)
73. Zhu, T., Li, H., Fu, Y.: Trim-sod: A multi-modal, multi-task, and multi-scale spacecraft optical dataset. Space: Science & Technology **5**, 0299 (2025)
74. Zhu, W., Cao, J., Xie, J., Yang, S., Pang, Y.: Clip-vis: Adapting clip for open-vocabulary video instance segmentation. IEEE TCSVT **35**(2), 1098–1110 (2024)