Unleashing the Temporal-Spatial Reasoning Capacity of GPT for Training-Free Audio and Language Referenced Video Object Segmentation

Shaofei Huang[1,2*], **Rui Ling**[3*], **Hongyu Li**[4*], **Tianrui Hui**[1†], **Zongheng Tang**[4], **Xiaoming Wei**[5], **Jizhong Han**[2], **Si Liu**[4†]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3]School of Computer Science and Engineering, Beihang University, Beijing, China
[4]School of Artificial Intelligence, Beihang University, Beijing, China
[5]Meituan, Beijing, China
{nowherespyfly, hongyuer0317, huitianrui}@gmail.com, weixiaoming@meituan.com,
{20373184, tzhhhh123, liusi}@buaa.edu.cn, hanjizhong@iie.ac.cn

Abstract

In this paper, we propose an Audio-Language-Referenced SAM 2 (AL-Ref-SAM 2) pipeline to explore the training-free paradigm for audio and language-referenced video object segmentation, namely AVS and RVOS tasks. The intuitive solution leverages GroundingDINO to identify the target object from a single frame and SAM 2 to segment the identified object throughout the video, which is less robust to spatiotemporal variations due to a lack of video context exploration. Thus, in our AL-Ref-SAM 2 pipeline, we propose a novel GPT-assisted Pivot Selection (GPT-PS) module to instruct GPT-4 to perform two-step temporal-spatial reasoning for sequentially selecting pivot frames and pivot boxes, thereby providing SAM 2 with a high-quality initial object prompt. Within GPT-PS, two task-specific Chain-of-Thought prompts are designed to unleash GPT's temporal-spatial reasoning capacity by guiding GPT to make selections based on a comprehensive understanding of video and reference information. Furthermore, we propose a Language-Binded Reference Unification (LBRU) module to convert audio signals into language-formatted references, thereby unifying the formats of AVS and RVOS tasks in the same pipeline. Extensive experiments show that our training-free AL-Ref-SAM 2 pipeline achieves performances comparable to or even better than fully-supervised fine-tuning methods.

Code — https://github.com/appletea233/AL-Ref-SAM2

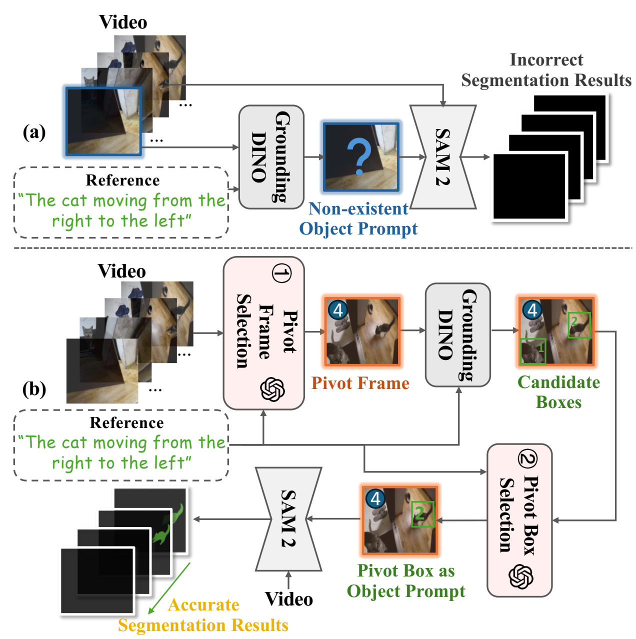

Figure 1: (a) For the training-free baseline, naively choosing the first frame to generate the object prompt may yield completely wrong predictions on the video since the first frame does not contain any relevant object. (b) Our method leverages GPT-4 to perform two-step temporal-spatial reasoning, selecting the frame and box that best reflects the reference information. The selected box serves as a more accurate object prompt to SAM 2 for better segmentation results.

Introduction

Video Object Segmentation (VOS), which involves segmenting and tracking specific objects throughout a video sequence, has garnered growing attention due to its significant potential in various real-world applications. Recent works further incorporate additional multimodal reference information, such as language and audio, making it more convenient and flexible to locate objects of interest in videos. For example, Referring Video Object Segmentation (RVOS) (Seo, Lee, and Han 2020) specifies the target object through language description, while Audio Visual Segmentation (AVS) (Zhou et al. 2022) segments objects emitting the sounds in the associated audio. However, most existing

*These authors contributed equally. Shaofei Huang is a visiting scholar at Hefei University of Technology.
†Tianrui Hui and Si Liu are the corresponding authors.

methods for these two tasks require extensive training in a fully-supervised manner, which suffers from high training costs and labor-intensive label annotations. Therefore, in this work, we propose to explore the training-free paradigm for multimodal information-aided VOS tasks.

Despite their cost-effectiveness and simplicity properties, training-free methods still exhibit significantly inferior performance compared with finetuning-based ones, owing to their lack of adaptability to the data distribution of specific VOS tasks. Recently, foundation models (Radford et al. 2021; Achiam et al. 2023; Liu et al. 2024b; Kirillov et al. 2023) have demonstrated strong generalization abilities across various tasks, illuminating a promising path to bridging this performance gap. For instance, SAM 2 (Ravi et al. 2024) shows powerful zero-shot transferability for promptable video segmentation, while GroundingDINO (Liu et al. 2023) excels at locating the target objects in a single image based on language descriptions. Leveraging these foundation models, an intuitive three-stage pipeline for training-free multimodal VOS tasks can be constructed (Ren et al. 2024): (1) extracting reference information from the multimodal input, (2) identifying the target object in the initial frame using GroundingDINO according to the extracted reference, and (3) segmenting the identified target object throughout the entire video using SAM 2. However, since GroundingDINO is designed for single-image grounding tasks, this intuitive solution may fail to identify target objects effectively due to its lack of video context exploration, making it less robust to spatiotemporal variations. As shown in Figure 1(a), if the target object is inaccurately identified or non-existent in the initial frame, subsequent frames may yield degraded segmentation results due to the incorrect object prompt to SAM 2.

To alleviate this limitation, we propose a novel two-step temporal-spatial reasoning flow that first selects a ***pivot frame*** from the entire video via temporal reasoning, and then identifies a ***pivot box*** from multiple candidate boxes on the pivot frame via spatial reasoning, thus providing SAM 2 with an accurate object prompt for initiating video segmentation. As shown in Figure 1(b), the pivot frame is defined as the specific frame where the target object (*i.e.*, the referent) clearly appears without being occluded or blurred, while the pivot box refers to the box on the pivot frame that best matches the reference information, *i.e.*, the referent's box. To implement this complex reasoning process within a training-free framework, we incorporate GPT-4 (Achiam et al. 2023) to leverage its vision-language comprehension and reasoning capacity. However, temporal and spatial reasoning are both inherently complex, necessitating an exhaustive comprehension of video and reference information. Naively providing GPT with abstract instructions such as *"select the frame where the target object appears clearly"* or *"locate the object according to the reference"* often fails to yield satisfactory results, as it typically chooses a middle frame as the pivot frame or the most salient object as the pivot box. We attribute this phenomenon to the logical shortcuts that GPT tends to take when confronted with complex reasoning tasks, often opting for the path of least resistance and resulting in superficial answers in the absence of explicit guidance (Wei et al. 2022). To tackle this issue, we meticulously design two task-specific Chain-of-Thought (CoT) prompts for the aforementioned two steps respectively, guiding GPT step-by-step to first form a comprehensive understanding of the video-reference pairs before answering the final questions. We implement the two steps as a GPT-assisted Pivot Selection (GPT-PS) module which unleashes the temporal-spatial reasoning capacity of GPT, prompting it to perform complex reasoning based on visual and reference information to make accurate judgments.

Additionally, given that audio signals are semantically ambiguous and intrinsically redundant, unifying audio into language format benefits more in the training-free setting for AVS task. To this end, we design a Language-Binded Reference Unification (LBRU) module to convert the audio signal into descriptions of the sounding objects from an acoustic perspective. Concretely, to eliminate the semantic ambiguity of audio signals and the interference of background noise, we incorporate visual context and leverage GPT-4 to identify the categories of the sounding objects from audio-video pairs. Symbolic representation is employed to encode video and audio data into sequences recognizable by GPT. By converting audio signals into higher-level, clearly defined language-formatted references, we can not only mitigate the inferior performance caused by the inherent shortcomings of audio, but also unify the formats of AVS and RVOS tasks, enabling our pipeline to handle both seamlessly. Integrating LBRU and GPT-PS modules, we name our pipeline as Audio-Language-Referenced SAM 2 (AL-Ref-SAM 2) which performs training-free unified audio- and language-referenced video object segmentation.

The contributions of our paper are summarized as follows: (1) We propose a GPT-assisted Pivot Selection (GPT-PS) module where GPT-4 is instructed to perform two-step temporal-spatial reasoning for selecting pivot frames and boxes that match with references, providing high-quality prompts to SAM 2 for precise video segmentation. (2) We propose a Language-Binded Reference Unification (LBRU) module that converts audio signals into language-formatted references, unifying AVS and RVOS tasks to be handled in the same pipeline. (3) Extensive experiments demonstrate that our training-free AL-Ref-SAM 2 pipeline achieves results that are comparable to, and in certain benchmarks even better than, those of finetuning-based methods on both tasks.

Related Works

Referring Video Object Segmentation

Different from Referring Image Segmentation (Hui et al. 2020; Liu et al. 2021; Huang et al. 2020; Liu, Jiang, and Ding 2024), Referring Video Object Segmentation (RVOS) (Seo, Lee, and Han 2020; Hui et al. 2021; Ding et al. 2022; Hui et al. 2023) aims to segment the target object matched with the description of a given sentence. ReferFormer (Wu et al. 2022) and MTTR (Botach, Zheltonozhskii, and Baskin 2022) utilize language as queries which attend to the relevant visual regions through Transformer models. Temporal information is also investigated to align object motion with language expression by reasoning across multi-

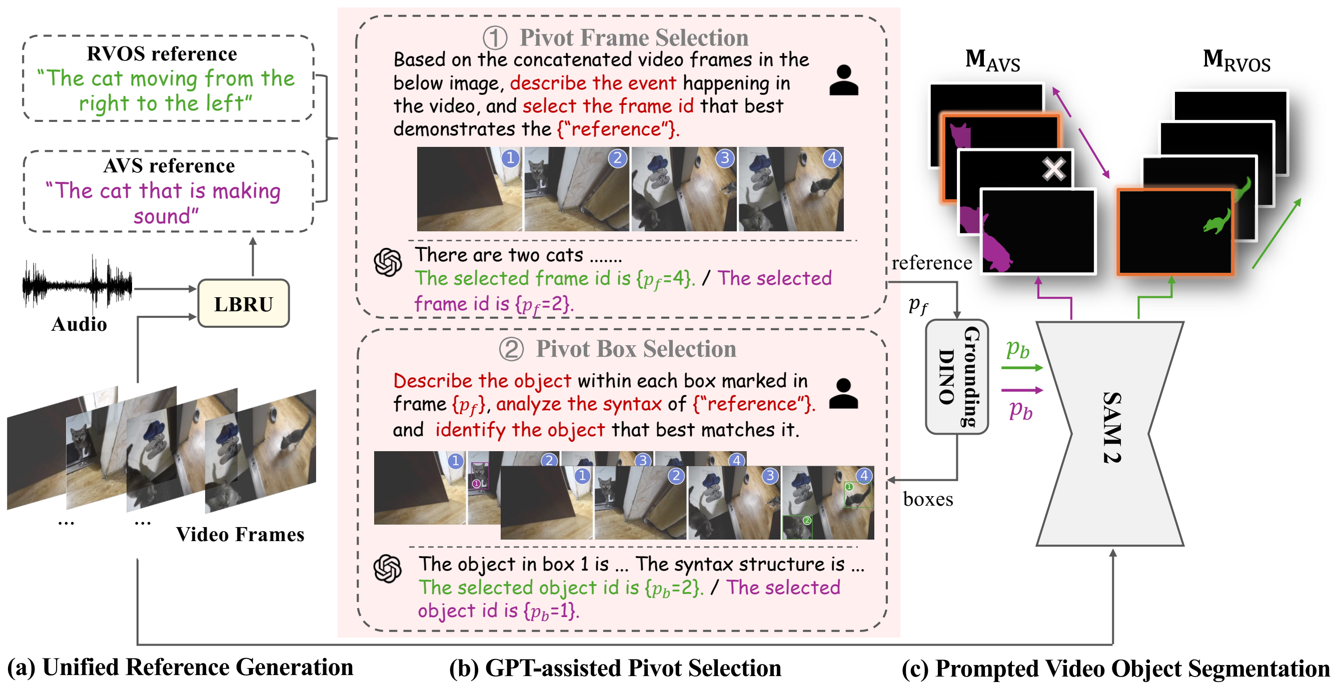

Figure 2: The overall pipeline of our proposed Audio-Language-Referenced SAM 2. (a) Generate a language-formatted reference that specifies the objects to be segmented for both RVOS and AVS tasks. (b) Select the pivot frame and pivot box through two-step temporal-spatial reasoning. (c) Prompt SAM 2 with the selected pivot box to obtain the mask sequence of the target object across the entire video. The symbol $\times$ on $\mathbf{M}_{\text{AVS}}$ represents the mask to be filtered out where the sound-emitting object is silent. *Different colors are used to denote the data flow of RVOS and AVS tasks respectively.*

ple temporal scales (Han et al. 2023) or temporal interaction between the global referent token and object queries (Tang, Zheng, and Yang 2023). LoSh (Yuan et al. 2024) proposes to utilize both long and short text expressions to understand the appearance and motion cues of the target object. In this paper, we tackle the RVOS task in the training-free setting by leveraging the temporal-spatial reasoning capacity of GPT.

Audio-Visual Segmentation

The goal of Audio-Visual Segmentation (Zhou et al. 2022) is to localize the objects emitting the given sounds in a video by pixel-level masks and predict their semantic category labels (Zhou et al. 2023) as well. AQFormer (Huang et al. 2023) constructs a set of object queries conditioned on audio information to associate visual regions of sounding objects with audio cues. (Chen et al. 2024) propose audio-visual supervised contrastive learning with an informative sample mining technique, aiming to utilize discriminative contrastive samples to strengthen cross-modal understanding. (Liu et al. 2024c) propose leveraging motion cues from neighboring frames and semantic cues from distant frames to extract information from abundant unlabeled frames for improving performance. In this paper, considering the inherent complexity and ambiguity of audio, we express it as a form of language and leverage foundation models to achieve unified and training-free video object segmentation.

Foundation Models

Foundation models are large, pre-trained models that serve as a general-purpose base for a wide range of downstream tasks across different domains (Shi et al. 2024; Wang et al. 2023). GPT-4 (Achiam et al. 2023) has garnered significant attention for its impressive conversational and reasoning capabilities. In addition, other open-source vision-language models (such as CLIP (Radford et al. 2021), LLaVA (Liu et al. 2024b), OPT (Zhang et al. 2022), Diffusion Models (Rombach et al. 2022; Zhang et al. 2023), *etc.*) have also demonstrated excellent multimodal understanding and generation abilities. In terms of visual foundation models, the well-known SAM (Kirillov et al. 2023), trained on over a billion masks, demonstrates powerful zero-shot image segmentation capabilities based on various prompts. SAM-Track (Cheng et al. 2023) and SAM 2 (Ravi et al. 2024) further extend the fine-grained perception ability of SAM to videos, enabling general segmentation and tracking of objects within video sequences. In this paper, we design a novel two-step temporal-spatial reasoning flow to exploit the strong task execution capability of foundation models for training-free multimodal video object segmentation.

Audio-Language-Referenced SAM 2

Pipeline Overview

The overall architecture of our Audio-Language-Referenced SAM 2 pipeline (AL-Ref-SAM 2) in illustrated in Figure 2.

Given a video clip consisting of T frames, and either a language description or an audio clip specifying the objects to be segmented (termed the *referent*), the goal of our AL-Ref-SAM 2 pipeline is to obtain the mask sequence $\mathbf{M}_{\mathrm{RVOS}}$ or $\mathbf{M}_{\mathrm{AVS}} \in \mathbb{R}^{T\times H\times W}$ of the referent across the whole video. Here, H and W denote the height and width of the video frames respectively. The language description is utilized in the RVOS task as the reference for indicating the referent. For the AVS task, we feed both the audio and video clips to the Language-Binded Reference Unification (LBRU) module to acquire a reference that describes the sounding objects from the acoustic aspect, thereby unifying the format of the AVS reference with that of the RVOS reference. Subsequently, the obtained reference and the video clip are processed by the GPT-assisted Pivot Selection (GPT-PS) module to identify a high-quality bounding box of the referent in a specific frame where the referent clearly appears. Finally, the selected bounding box serves as the pivot box to effectively prompt SAM 2 to segment the referent and propagate its mask forward and backward through the entire video clip. For the AVS task, we use sound event detection to segment the audio clip and filter out silent frames.

Language-Binded Reference Unification

AVS aims to segment the sounding objects within the video clip based on its corresponding audio content, which can also be formulated as using the reference of "*the* [OBJ] *that is making sound*" for segmentation, where "[OBJ]" denotes the category of the specific sounding object in the video. In this way, the audio clip in the AVS task can be converted into a language-formatted reference for facilitating better task unification. To obtain the categories of all the sounding objects, an intuitive approach involves applying an audio classifier (*e.g.*, BEATs (Chen et al. 2022)) to the audio clip, classifying it into several categories. However, due to the presence of background noise and the ambiguity of audio information, this audio-only approach may collect incorrect or unnecessary object categories, leading to suboptimal segmentation performance. Therefore, we incorporate visual context and leverage the inherent vision-language understanding capabilities of MLLMs (*e.g.*, GPT-4) to accurately identify the categories of the actual sounding objects present in the video.

Given that GPT-4 is currently unable to comprehend audio and video data, we first encode the audio and video data into GPT-recognizable sequences through symbolic representation. Specifically, as shown in Figure 3, for the audio clip, we feed it into the audio classifier and retain the categories with the top-k classification confidence scores. The text of the retained categories is then organized into a list as $\mathbf{X}_{\mathrm{a}} = [\mathrm{CLS}_1, \mathrm{CLS}_2, ..., \mathrm{CLS}_k]$. As for the video clip, we evenly sample m frames and concatenate them sequentially into a single image, with the frame ID marked on each frame. The concatenated image is denoted as $\mathbf{X}_{\mathrm{v}}$. We then combine the above symbolic representations with our carefully curated language command $\mathbf{X}_{\mathrm{l}}$ as the prompt $\mathbf{X}_{\mathrm{pa}}$ to GPT: $\mathbf{X}_{\mathrm{pa}} = [\mathbf{X}_{\mathrm{l}}, \mathbf{X}_{\mathrm{a}}, \mathbf{X}_{\mathrm{v}}]$. The goal of $\mathbf{X}_{\mathrm{pa}}$ is to guide GPT step by step in filtering and merging audio labels based on the content of the video frames, ultimately outputting the

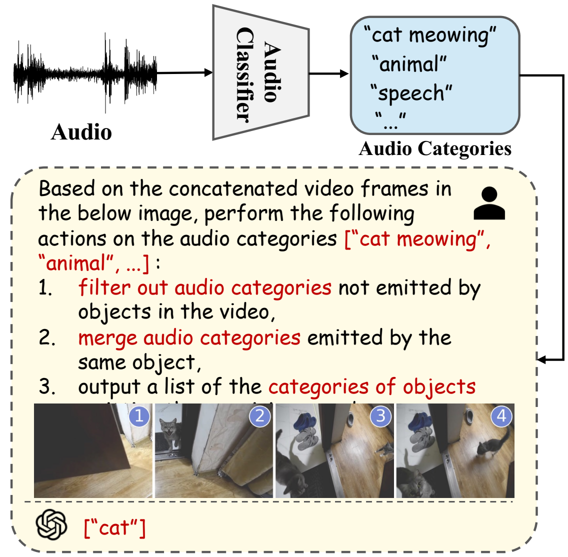

Figure 3: Detailed illustration of our Language-Binded Reference Unification module.

categories of objects emitting the remaining sounds. These categories are denoted as $[\mathrm{OBJ}_1, \mathrm{OBJ}_2..., \mathrm{OBJ}_n]$, where n represents the total number of sounding object categories in the video. We convert each category to a separate reference and process it individually.

GPT-Assisted Pivot-Selection

Our GPT-assisted Pivot Selection (GPT-PS) module aims to leverage GPT's vision-language reasoning capabilities to obtain a highly distinguishable bounding box of the referent, thus prompting SAM 2 for precise segmentation. The reasoning process involves two steps. In the first step, temporal reasoning is conducted to select a pivot frame from the entire video, in which the referent can be clearly distinguished. Subsequently, spatial reasoning is carried out on the pivot frame to select the pivot box that best matches the reference description from multiple candidates within this frame. Utilizing the pivot frame as the starting point for SAM 2's segmentation process, we propagate the object mask within the pivot box forward and backward throughout the entire video to obtain the complete mask sequence of the referent.

Step 1: Pivot Frame Selection. As illustrated in the upper part of Figure 2(b), we employ the same way as described in LBRU to obtain the symbolic representation $\mathbf{X}_{\mathrm{v}}$ of video information. The colored frame numbers marked on $\mathbf{X}_{\mathrm{v}}$ can inform GPT of the order of the concatenated frames, thereby enhancing its comprehension of the video information. Pivot frame selection is a complex temporal reasoning process that necessitates GPT to both comprehend the events occurring in the video and preliminarily identify the actual object mentioned in the reference. To this end, we design the Pivot-Frame CoT prompt, $\mathbf{X}_{\mathrm{pf}}$, which instructs GPT to first describe the temporal event for enhancing its comprehensive understanding of the video content, and then select the pivot frame from the sampled frames in $\mathbf{X}_{\mathrm{v}}$ where the reference

Method	Reference	Ref-YouTube-VOS $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	Ref-DAVIS17 $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	MeViS $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Fully-Supervised Fine-Tuning										
ReferFormer (2022)	[CVPR'22]	62.9	61.3	64.6	61.1	58.1	64.1	31.0	29.8	32.2
OnlineRefer (2023)	[ICCV'23]	62.9	61.0	64.7	62.4	59.1	65.6	-	-	-
HTML (2023)	[ICCV'23]	63.4	61.5	65.2	62.1	59.2	65.1	-	-	-
SgMg (2023)	[ICCV'23]	65.7	63.9	67.4	63.3	60.6	66.0	-	-	-
LMPM (2023)	[ICCV'23]	-	-	-	-	-	-	37.2	34.2	40.2
TempCD (2023)	[ICCV'23]	65.8	63.6	68.0	64.6	61.6	67.6	-	-	-
SOC (2024)	[NIPS'23]	66.0	64.1	67.9	64.2	61.0	67.4	-	-	-
VD-IT (2024)	[ECCV'24]	66.5	64.4	68.5	63.0	59.9	66.1	-	-	-
LoSh (2024)	[CVPR'24]	67.2	65.4	69.0	64.3	61.8	66.8	-	-	-
DsHmp (2024)	[CVPR'24]	67.1	65.0	69.1	64.9	61.7	68.1	46.4	43.0	49.8
Weakly-Supervised Fine-Tuning										
WRVOS (2023)	[arXiv'23]	46.6	45.6	47.6	47.3	44.6	50.0	-	-	-
GroPrompt (2024)	[CVPRW'24]	65.5	64.1	66.9	70.6	67.8	73.3	-	-	-
Training-Free										
G-L + SAM 2 (2023)	[CVPR'23]	27.0	24.3	29.7	40.6	37.6	43.6	23.7	20.4	30.0
G-L (SAM) + SAM 2 (2023)	[CVPR'23]	33.6	29.9	37.3	46.9	44.0	49.7	26.6	22.7	30.5
Grounded-SAM (2024)*	[arXiv'24]	62.3	61.0	63.6	65.2	62.3	68.0	-	-	-
Grounded-SAM 2 (2024)†	[arXiv'24]	64.8	62.5	67.0	66.2	62.6	69.7	38.9	35.7	42.1
AL-Ref-SAM 2 (Ours)	-	**67.9**	**65.9**	**69.9**	**74.2**	**70.4**	**78.0**	**42.8**	**39.5**	**46.2**

Table 1: Comparison with state-of-the-art methods on the validation sets of Ref-YouTube-VOS, Ref-DAVIS17 and MeViS datasets. * Results are adopted from GroPrompt. † Results are obtained by running the official code.

Method	Reference	S4 $\mathcal{M}_\mathcal{J}$	S4 $\mathcal{M}_\mathcal{F}$	MS3 $\mathcal{M}_\mathcal{J}$	MS3 $\mathcal{M}_\mathcal{F}$	AVSS $\mathcal{M}_\mathcal{J}$	AVSS $\mathcal{M}_\mathcal{F}$	AVSS-V2-Binary $\mathcal{M}_\mathcal{J}$	AVSS-V2-Binary $\mathcal{M}_\mathcal{F}$
Fully-Supervised Fine-Tuning									
TPAVI (2022; 2023)	[ECCV'22]	78.7	87.9	54.0	64.5	29.8	35.2	62.5	75.6
AQFormer (2023)	[IJCAI'23]	81.6	89.4	61.1	72.1	-	-	-	-
CATR (2023)	[ACM MM'23]	81.4	89.6	59.0	70.0	32.8	38.5	-	-
BAVS (2024a)	[TMM'24]	82.0	88.6	58.6	65.5	32.6	36.4	-	-
Audio-SAM (2024)	[AAAI'24]	56.3	72.7	33.7	45.9	-	-	57.4	68.4
SAM-Fusion (2024)	[AAAI'24]	71.9	77.5	50.6	63.7	-	-	60.2	72.4
GAVS (2024)	[AAAI'24]	80.1	90.2	63.7	77.4	-	-	67.7	78.8
AVSegFormer (2024)	[AAAI'24]	82.1	89.9	58.4	69.3	36.7	42.0	64.3	75.9
COMBO (2024)	[CVPR'24]	84.7	91.9	59.2	71.2	42.1	46.1	-	-
Weakly-Supervised Fine-Tuning									
WS-AVS (2024)	[NeurIPS'24]	34.1	51.8	30.9	46.9	-	-	-	-
MoCA (2024)	[arXiv'24]	68.0	79.0	57.0	62.0	31.0	33.0	-	-
Training-Free									
AT-GDINO-SAM (2024)	[arXiv'24]	38.0	46.0	25.0	29.0	24.0	25.0	-	-
SAM-BIND (2024)	[arXiv'24]	42.0	51.0	28.0	36.0	24.0	26.0	-	-
OWOD-BIND (2024)	[arXiv'24]	58.0	67.0	34.0	44.0	26.0	29.0	-	-
AL-Ref-SAM 2 (Ours)	-	**70.5**	**81.1**	**48.6**	**53.5**	**36.0**	**39.8**	**59.2**	**66.2**

Table 2: Comparison with state-of-the-art methods on the different subsets of the AVSBench dataset. ".0" in the results of other training-free methods is due to the rounding errors from their original papers.

content is most easily recognized. The obtained frame ID is denoted as p_f.

Step 2: Pivot Box Selection. As illustrated in the lower part of Figure 2(b), we first employ GroundingDINO to predict candidate boxes of the referent on the pivot frame according to the reference description. Since GroundingDINO is designed for grounding tasks on a single image, whereas our reference contains extensive temporal-related information, it is challenging to accurately predict the target object based solely on a single frame. Thus, we lower the confidence threshold for GroundingDINO to generate multiple candidates, maximizing the likelihood of the inclusion of the actual referent. Subsequently, we paint the candidate boxes on the pivot frame and mark each with a box ID to designate the regions for GPT's attention. We also sequentially concatenate the other sampled frames with the marked pivot frame to provide temporal context for distinguishing the referent from multiple candidates. To further aid this process, a Pivot-Box CoT prompt, $\mathbf{X}_{\mathrm{pb}}$, is designed to guide GPT step-by-step in reasoning about the pivot box that best matches the reference description on the pivot frame. Specifically, based on the prior knowledge of the video event obtained in step 1, $\mathbf{X}_{\mathrm{pb}}$ first instructs GPT to describe the objects in each box, enabling it to comprehend candidates' appearance, motion, and interrelationships. Afterward, GPT is required to perform syntactic analysis of the reference description to accurately identify the subject being referred to. Finally, the candidate box that best corresponds to the description of the identified subject is selected as the pivot box p_b. For illustrative purposes, we present an abridged version of the prompts in Figure 2(b).

Experiments

Datasets and Evaluation Metrics

We adopt Ref-YouTube-VOS (Seo, Lee, and Han 2020), Ref-DAVIS17 (Khoreva, Rohrbach, and Schiele 2019), MeViS (Ding et al. 2023) for RVOS evaluation, and AVSBench (Zhou et al. 2022) datasets for AVS evaluation respectively. In terms of evaluation metrics, we adopt region similarity $\mathcal{J}$ (average IoU), contour accuracy $\mathcal{F}$ and their average $\mathcal{J}\&\mathcal{F}$ for RVOS. The metrics $\mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{F}}$ for AVS are the same as $\mathcal{J}$ and $\mathcal{F}$.

Implementation Details

We adopt the `sam2_hiera_large` version of SAM 2 as the video segmentor and `swinb_cogcoor` version of GroundingDINO as the grounding model. For the RVOS task, we first divide the input video into several clips based on the sampling number and interval of frames. Then, we conduct pivot frame and box selection within each clip. All the obtained pivot boxes are used as prompts to SAM 2 while the pivot frame in the middle clip is used as the starting frame of mask propagation. In each clip, the number of sampled frames is 5 for all datasets while the sampling interval between frames varies across different datasets. For the Ref-YouTube-VOS dataset, which already employs a 5-frame sampling interval, we set an additional sampling interval of 2 frames, resulting in an actual sampling interval of 10 frames. For the Ref-DAVIS17 and MeViS datasets, the sampling interval is set to 5 frames. We set GroundingDINO's `text_threshold` to 0.2 and `box_threshold` to 0.15.

For the AVS task, we adopt BEATs (Chen et al. 2022) as the audio classifier and keep audio categories with the top-5 confidence scores in the LBRU module. Since the video length is relatively shorter in the AVSBench dataset, we do not divide its videos into clips. For each video in AVSBench, we sample 5 frames for the S4 and MS3 settings and 10 frames for the AVSS setting. We set GroundingDINO's `text_threshold` to 0.25 and `box_threshold` to 0.25.

Comparison with State-of-the-Art Methods

Table 1 and Table 2 present the quantitative comparison results between our method and previous state-of-the-art methods on RVOS and AVS tasks, respectively. "G-L + SAM 2" denotes adopting a zero-shot referring image segmentation method Global-Local (Yu, Seo, and Son 2023) to obtain the referent mask in the first frame and then using it as the prompt to SAM 2. As observed in Table 1, our proposed training-free AL-Ref-SAM 2 achieves the best performance in both training-free and weakly-supervised fine-tuning settings, and it also outperforms most fully-supervised fine-tuning methods on three datasets. Furthermore, Table 2 shows that our method significantly reduces the performance gap between training-free methods and those fine-tuned with full supervision in the AVS task. These results demonstrate that by designing effective chain-of-thought and unification strategies, we can fully unleash the reasoning and perception capabilities of foundation models, enabling a unified pipeline to empower multiple specific VOS tasks.

Ablation Studies

We conduct ablation studies on the RVOS and AVS tasks to verify the effectiveness of different designs in our pipeline.

	Ref-YouTube-VOS			Ref-DAVIS17		
Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
G-DINO + SAM 2	64.8	62.5	67.0	66.2	62.6	69.7
+ PF Select	65.6	63.6	67.6	71.1	67.0	74.8
+ PF & PB Select	**67.9**	**65.9**	**69.9**	**74.2**	**70.4**	**78.0**

Table 3: Ablation study of the two-step reasoning in GPT-assisted pivot selection.

Two-step reasoning in GPT-PS. As shown in Table 3, we validate the effectiveness of the two-step pivot selection process proposed in our GPT-PS on two RVOS datasets. In the first row, "G-DINO + SAM 2" represents our baseline method, which directly uses GroundingDINO on the first frame of the video to obtain the referent's box prediction based on the input language. This box is then used as a prompt for SAM 2 to obtain the video segmentation results. In the second row, "+ PF Select" denotes integrating the first step of selecting the pivot frame in our GPT-PS into the baseline model. The last row denotes further integrating the second step of selecting the pivot box upon the first step. The results show that using only the first step, as well as employing both steps of our proposed GPT-PS, consistently improves the model's performance. This indicates that leveraging GPT's temporal-spatial reasoning ability to obtain higher-quality prompts can better tap into SAM 2's potential. Similarly, results in Table 4 also witness a consistent improvement of both steps in GPT-PS for the AVS task. This further demonstrates that our proposed temporal-spatial reasoning strategy has excellent generalization ability for various multimodal-referenced VOS tasks under the training-free setting.

Method	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
BEATs Reference	53.7	59.6
LBRU Reference	56.9	64.5
+ PF Selection	58.4	65.9
+ PF & PB Selection	**59.2**	**66.2**

Table 4: Component ablation on AVSS-V2-Binary dataset.

LBRU reference. Furthermore, we also verify the effectiveness of our proposed LBRU on the AVSS-V2-Binary dataset, as shown in the first two rows of Table 4. "BEATs

(a) *"A bike ridden behind another bike"*

(b) *"A wet black seal sitting in the bottom left foreground"*

Figure 4: Qualitative comparison between our method and the baseline GD-SAM 2 on the Ref-YouTube-VOS dataset.

Reference" denotes that the reference to be fed into GroundingDINO + SAM 2 is generated with the audio categories predicted by BEATs only, which is suboptimal. Incorporating our proposed LBRU to extract reference achieves notable performance improvement, showing that integrating visual context and using GPT to convert audio signals to language formats can express the audio context more precisely.

Method	Ref-YouTube-VOS $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
First Frame	64.8	62.5	67.0
Middle Frame	64.9	62.7	67.1
Last Frame	58.5	56.5	60.5
PF Selection	**65.6**	**63.6**	**67.6**

Table 5: Results of different frame selection strategies.

Different frame selection strategies. In Table 5, we compare the results of different frame selection strategies on the Ref-YouTube-VOS dataset. The first three rows select three different fixed frames as the pivot frame. Our proposed GPT-assisted pivot frame selection achieves the best performance among the four strategies, showing that leveraging GPT's temporal reasoning ability allows us to select a pivot frame that better reflects the reference information, thereby establishing a strong visual context for the entire segmentation.

Method	Ref-YouTube-VOS $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
w/o GPT Assisted Box Selection	65.6	63.6	67.6
Select Referent *w/o* Description	63.3	61.0	65.5
Describe Then Select	65.6	63.4	67.8
w/ Syntax Analysis of Reference	**67.9**	**65.9**	**69.9**

Table 6: Different prompts to GPT for pivot box selection.

Different box selection prompts to GPT. In Table 6, we present the results of using different prompts to GPT in the pivot box selection step. The first row means we naively adopt the highest-scoring box predicted by GroundingDINO on the pivot frame as the pivot box. The second row indicates that we directly instruct GPT to select the object box that matches the reference information, showing that providing GPT with naive instructions makes it even harder to perform correct reasoning. The third row represents that we first have GPT provide a detailed description of the scene and objects before selecting the pivot box. The last row further includes syntax analysis of the language reference, resulting in significant improvement. These results show that guiding GPT to conduct reasoning with our complete PF CoT prompt benefits the accurate pivot box selection.

Qualitative Results

In Figure 4, we compare the qualitative results of our method with those of the baseline (GroundingDINO + SAM 2) on the Ref-YouTube-VOS dataset. These results demonstrate that our method has a stronger ability to distinguish between objects of the same category compared to the baseline. In the first row, even though only a small part of the bicycle's handlebar is visible, our method still accurately locates the referred bicycle in the back based on cues like "behind".

Conclusion

In this paper, we propose a Audio-Language-Referenced SAM 2 (AL-Ref-SAM 2) pipeline for training-free audio and language-referenced video object segmentation, namely AVS and RVOS tasks. A novel GPT-assisted Pivot Selection (GPT-PS) module is developed to guide GPT-4 for selecting pivot frames and boxes with two-step temporal-spatial reasoning, thereby providing SAM 2 with high-quality initial prompts. Two task-specific Chain-of-Thought prompts are designed for explicit guidance of GPT within GPT-PS. Additionally, a Language-Binded Reference Unification (LBRU) module which converts audio signals into language-formatted references is also designed, thus unifying AVS and RVOS tasks within the same pipeline. Extensive experiments on both tasks demonstrate that our training-free AL-Ref-SAM 2 pipeline achieves results comparable to, or even better than, those of finetuning-based methods.

Acknowledgments

This research was supported in part by National Science and Technology Major Project (2022ZD0115502), National Natural Science Foundation of China (No. U23B2010, No. 62461160308), Zhejiang Provincial Natural Science Foundation of China (Grant No. LDT23F02022F02), Beijing Natural Science Foundation (No. L231011), Beihang World TOP University Cooperation Program, and Meituan.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bhosale, S.; Yang, H.; Kanojia, D.; Deng, J.; and Zhu, X. 2024. Unsupervised Audio-Visual Segmentation with Modality Alignment. *arXiv preprint arXiv:2403.14203*.

Botach, A.; Zheltonozhskii, E.; and Baskin, C. 2022. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*.

Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; and Wei, F. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Chen, Y.; Liu, Y.; Wang, H.; Liu, F.; Wang, C.; Frazer, H.; and Carneiro, G. 2024. Unraveling Instance Associations: A Closer Look for Audio-Visual Segmentation. In *CVPR*.

Cheng, Y.; Li, L.; Xu, Y.; Li, X.; Yang, Z.; Wang, W.; and Yang, Y. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558*.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*.

Ding, Z.; Hui, T.; Huang, J.; Wei, X.; Han, J.; and Liu, S. 2022. Language-bridged spatial-temporal interaction for referring video object segmentation. In *CVPR*.

Gao, S.; Chen, Z.; Chen, G.; Wang, W.; and Lu, T. 2024. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*.

Han, M.; Wang, Y.; Li, Z.; Yao, L.; Chang, X.; and Qiao, Y. 2023. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *ICCV*.

He, S.; and Ding, H. 2024. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*.

Huang, S.; Hui, T.; Liu, S.; Li, G.; Wei, Y.; Han, J.; Liu, L.; and Li, B. 2020. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*.

Huang, S.; Li, H.; Wang, Y.; Zhu, H.; Dai, J.; Han, J.; Rong, W.; and Liu, S. 2023. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*.

Hui, T.; Huang, S.; Liu, S.; Ding, Z.; Li, G.; Wang, W.; Han, J.; and Wang, F. 2021. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *CVPR*.

Hui, T.; Liu, S.; Ding, Z.; Huang, S.; Li, G.; Wang, W.; Liu, L.; and Han, J. 2023. Language-aware spatial-temporal collaboration for referring video segmentation. *TPAMI*.

Hui, T.; Liu, S.; Huang, S.; Li, G.; Yu, S.; Zhang, F.; and Han, J. 2020. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*.

Khoreva, A.; Rohrbach, A.; and Schiele, B. 2019. Video object segmentation with language referring expressions. In *ACCV*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*.

Li, K.; Yang, Z.; Chen, L.; Yang, Y.; and Xiao, J. 2023. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM MM*.

Lin, C.-S.; Liu, I.; Chen, M.-H.; Wang, C.-Y.; Liu, S.; Wang, Y.-C. F.; et al. 2024. GroPrompt: Efficient Grounded Prompting and Adaptation for Referring Video Object Segmentation. *arXiv preprint arXiv:2406.12834*.

Liu, C.; Jiang, X.; and Ding, H. 2024. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*.

Liu, C.; Li, P.; Zhang, H.; Li, L.; Huang, Z.; Wang, D.; and Yu, X. 2024a. BAVS: bootstrapping audio-visual segmentation by integrating foundation knowledge. *TMM*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *NeurIPS*.

Liu, J.; Liu, Y.; Zhang, F.; Ju, C.; Zhang, Y.; and Wang, Y. 2024c. Audio-Visual Segmentation via Unlabeled Frame Exploitation. In *CVPR*.

Liu, S.; Hui, T.; Huang, S.; Wei, Y.; Li, B.; and Li, G. 2021. Cross-modal progressive comprehension for referring segmentation. *TPAMI*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Luo, Z.; Xiao, Y.; Liu, Y.; Li, S.; Wang, Y.; Tang, Y.; Li, X.; and Yang, Y. 2024. Soc: Semantic-assisted object cluster for referring video object segmentation. *NeurIPS*.

Miao, B.; Bennamoun, M.; Gao, Y.; and Mian, A. 2023. Spectrum-guided multi-granularity referring video object segmentation. In *ICCV*.

Mo, S.; and Raj, B. 2024. Weakly-supervised audio-visual segmentation. *NeurIPS*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Seo, S.; Lee, J.-Y.; and Han, B. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*.

Shi, H.; Han, S.; Huang, S.; Liao, Y.; Li, G.; Kong, X.; Zhu, H.; Wang, X.; and Liu, S. 2024. Mask-enhanced segment anything model for tumor lesion semantic segmentation. In *MICCAI*.

Tang, J.; Zheng, G.; and Yang, S. 2023. Temporal collection and distribution for referring video object segmentation. In *ICCV*.

Wang, Y.; Huang, S.; Gao, Y.; Wang, Z.; Wang, R.; Sheng, K.; Zhang, B.; and Liu, S. 2023. Transferring CLIP's Knowledge into Zero-Shot Point Cloud Semantic Segmentation. In *ACM MM*.

Wang, Y.; Liu, W.; Li, G.; Ding, J.; Hu, D.; and Li, X. 2024. Prompting segmentation with sound is generalizable audio-visual source localizer. In *AAAI*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Wu, D.; Wang, T.; Zhang, Y.; Zhang, X.; and Shen, J. 2023. Onlinerefer: A simple online baseline for referring video object segmentation. In *ICCV*.

Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022. Language as queries for referring video object segmentation. In *CVPR*.

Yang, Q.; Nie, X.; Li, T.; Gao, P.; Guo, Y.; Zhen, C.; Yan, P.; and Xiang, S. 2024. Cooperation Does Matter: Exploring Multi-Order Bilateral Relations for Audio-Visual Segmentation. In *CVPR*.

Yu, S.; Seo, P. H.; and Son, J. 2023. Zero-shot referring image segmentation with global-local context features. In *CVPR*.

Yuan, L.; Shi, M.; Yue, Z.; and Chen, Q. 2024. Losh: Long-short text joint prediction network for referring video object segmentation. In *CVPR*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.

Zhao, W.; Nan, K.; Zhang, S.; Chen, K.; Lin, D.; and You, Y. 2023. Learning referring video object segmentation from weak annotation. *arXiv preprint arXiv:2308.02162*.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2023. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–visual segmentation. In *ECCV*.

Zhu, Z.; Feng, X.; Chen, D.; Yuan, J.; Qiao, C.; and Hua, G. 2024. Exploring Pre-trained Text-to-Video Diffusion Models for Referring Video Object Segmentation. *arXiv preprint arXiv:2403.12042*.