Referring and Reasoning Camouflaged Object Segmentation in Audio-Visual Scenes

Tianxin Han[1,2] **Qing Dong**[1*] **Xingwei Wang**[1] **Jie Jia**[1]

[1]Northeastern University, Shenyang, China
[2]Ghent University–imec, Ghent, Belgium

tianxinhan797@gmail.com

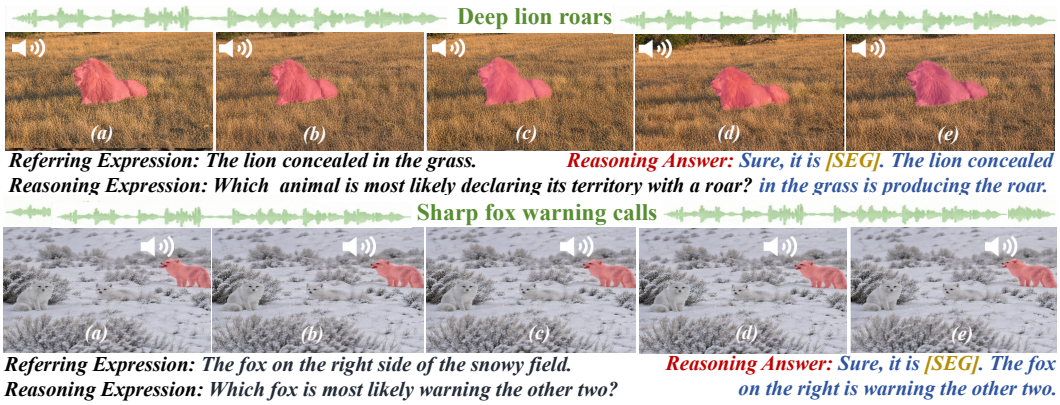

Figure 1: Examples of the proposed R2-AVCOS benchmark. Compared with existing COS benchmarks, R2-AVCOS has three key characteristics: 1) supporting both referring and reasoning expressions for audio-visual camouflaged target specification; 2) emphasizing audio content understanding beyond simple sound presence; 3) incorporating complex reasoning and world knowledge, requiring models to explain their segmentation decisions. These characteristics make R2-AVCOS closer to real-world camouflaged perception scenarios.

Abstract

Camouflaged Object Segmentation (COS) aims to identify objects visually hidden in their surrounding environments. Existing COS benchmarks mainly focus on image-level or video-level visual perception, while real-world camouflaged scenes are often accompanied by audio signals and user intentions, where sound semantics and textual expressions provide critical information for localizing hidden targets. To extend the boundary of COS, we introduce a new task, termed Referring and Reasoning Audio-Visual Camouflaged Object Segmentation (R2-AVCOS), which aims to segment camouflaged targets in audio-visual scenes according to textual expressions with referring or reasoning intentions. This task emphasizes understanding audio content and incorporates complex reasoning and world knowledge into expressions. To support this task, we construct R2-AVCOSBench, the first audio-visual benchmark with pixel-level annotations for camouflaged objects specified by referring expressions or inferred through reasoning. It contains 2,654 audio-visual camouflaged videos, 21,232 annotated frames, and 32,329 expressions, including 17,543 referring and 14,786 reasoning expressions. Furthermore, we

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

propose Camouflaged Instructed Segmentation Assistant (CISA), a baseline model built upon a Multimodal Large Language Model (MLLM). CISA understands complex textual and audio-visual cues and performs referring- and reasoning-based camouflaged object segmentation. Extensive experiments show that CISA achieves strong referring and reasoning segmentation ability in audio-visual camouflaged scenes and obtains competitive results on related tasks.

1 Introduction

In real-world scenarios, visual information is often accompanied by multimodal information such as audio and text, which provide valuable spatiotemporal and semantic cues for locating objects of interest [1]. For machine perception systems, this ability is particularly important in natural environments, where many targets resemble their surroundings and are difficult to recognize or segment from visual appearance alone. For example, a frog hidden in grass may be perceived mainly through its calls, while a bird on a tree trunk may require sound direction, local texture, and scene semantics for localization. In such cases, a single visual modality is often insufficient to support the expected segmentation behavior. As a representative visual perception task, Camouflaged Object Segmentation (COS) aims to delineate targets concealed within surrounding scenes [2], and is crucial for medical image analysis, surface defect inspection, and pest monitoring [3]. Existing studies have explored COS under relatively limited settings [4]. Image-level COS focuses on static background-similar targets but ignores temporal dynamics and audio information. Video-level COS introduces motion and temporal modeling, yet still mainly relies on visual appearance [5]. Text-guided COS incorporates language descriptions, but usually relies on explicit textual attributes or categories and lacks audio understanding [4]. Audio-Visual Segmentation (AVS) uses sound to segment sounding objects, but mainly targets common visible objects, lacking modeling of camouflaged targets and the ability to specify particular targets according to user intent [6].

Recently, Referring Audio-Visual Segmentation (Ref-AVS) has used natural language to specify user-interested targets in audio-visual scenes [7]. However, existing methods are still mainly designed for common visible objects, and the text-audio association often remains shallow, focusing on sound presence, intensity, or temporal order [8]. This is insufficient for audio-visual camouflaged scenes, where models must jointly understand sound semantics, scene context, and commonsense knowledge to identify hidden targets. For example, "which hidden animal is warning its companions?" cannot be solved by simply detecting the sounding region; the model must understand the meaning of the sound and connect it with weak camouflaged visual patterns. Similarly, "the hidden animal most likely engaging in courtship" requires understanding sound content, behavioral meaning, and natural-scene knowledge to determine the target and its pixel-level mask. Therefore, existing methods have not fully addressed COS in dynamic audio-visual scenes. Traditional COS lacks audio-language interaction, while existing AVS and Ref-AVS methods lack modeling of camouflaged targets, complex reasoning, and world knowledge. In real-world scenarios, users may describe their intent indirectly through sound, behavior, temporal relations, or scene semantics rather than target categories or appearance details, requiring models to perform both referring and reasoning-based audio-visual segmentation.

To this end, we propose **Referring and Reasoning Audio-Visual Camouflaged Object Segmentation (R2-AVCOS)**, which segments camouflaged targets in audio-visual scenes according to textual expressions, as shown in Fig. 1 and Fig. 2. It includes two expression types: referring expressions that specify targets through audio, visual, and temporal information, and reasoning expressions that identify targets through sound semantics, camouflaged visual patterns, scene context, and commonsense knowledge. To support this task, we construct **R2-AVCOSBench**, the first benchmark for this setting, containing 2,654 audio-visual camouflaged videos, 21,232 annotated frames, and 32,329 expressions, including 17,543 referring and 14,786 reasoning expressions. Unlike existing COS datasets that mainly focus on visual perception, R2-AVCOSBench introduces audio and textual intent. Compared with AVS datasets, it emphasizes audio-content understanding rather than sound presence, and further incorporates complex reasoning and world knowledge for fine-grained audio-visual camouflaged perception.

We futher introduce a multimodal large language model based baseline, **Camouflaged Instructed Segmentation Assistant (CISA)**. This model integrates textual, audio, and visual inputs, performs COS according to referring or reasoning expressions, and provides prediction explanations for complex reasoning samples. We adopt an audio-visual interleaving strategy to achieve parameter-free temporal

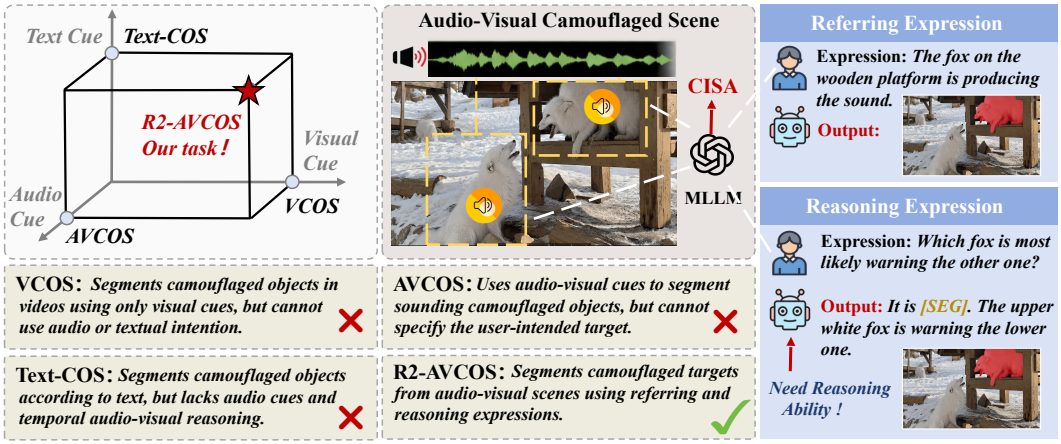

Figure 2: Illustration of the proposed **R2-AVCOS** task. Compared with other related tasks, our task jointly uses visual, audio, and textual cues to segment camouflaged targets specified by referring or reasoning expressions, requiring fine-grained audio-visual understanding and reasoning ability.

alignment between audio and video frames for efficient and accurate temporal segmentation. Experimental results show that the proposed method outperforms existing methods on R2-AVCOSBench and demonstrates favorable generalization ability on related tasks and datasets. In summary, our contributions are: **1)** We propose R2-AVCOS, a new task that extends COS from visual perception to audio-visual scenes with textual intent and complex reasoning. **2)** We construct R2-AVCOSBench, a large-scale benchmark with textual expressions, pixel-level masks, and explanation annotations, emphasizing joint audio, visual, and contextual understanding. **3)** We introduce referring and reasoning expressions to improve dataset interactivity and challenge, requiring models to identify camouflaged targets from user intent, sound semantics, temporal relations, and world knowledge. **4)** We develop an MLLM-based baseline that uniformly handles textual referring, audio-visual understanding, and reasoning segmentation, with extensive experiments validating its effectiveness.

2 Related Work

Camouflaged Object Segmentation. COS aims to segment targets visually concealed in their surroundings. Representative benchmarks have promoted COS research, but they still mainly rely on static visual information and provide only pixel-level masks, making it difficult to evaluate intent reasoning or multimodal instruction following [2, 9]. Video COS exploits motion cues and temporal consistency for dynamic scenes, but still depends mainly on visual information and overlooks audio signals commonly available in real-world videos [10, 11, 5]. Recent text-based COS studies [12, 13, 4] introduce language prompts, but their expressions usually specify categories or attributes directly, reducing the task to prompt-conditioned mask prediction. Therefore, existing COS, video COS, and text-COS settings remain insufficient for target understanding in realistic audio-visual camouflaged scenes, where users may specify targets through referring or reasoning expressions. Models thus need to jointly understand audio semantics, visual camouflage cues, temporal information, scene context, and commonsense knowledge to identify and segment the user-intended target.

Audio-Visual Scene Perception. Audio-visual scene perception integrates auditory and visual information for scene understanding, simulating human multimodal perception [6, 14]. Compared with single-modality perception, audio-visual learning is more effective when visual information is incomplete, ambiguous, or cluttered. However, datasets containing audio, visual, and language information remain limited. Music-AVQA [15] and AVQA [16] focus on audio-visual question answering, while Ref-AVS [7] studies referring audio-visual segmentation. Ying et al. [8] further explored referring audio-visual segmentation with full-modality expressions. Nevertheless, existing tasks are not designed for visually camouflaged scenes and rarely emphasize reasoning over concealed sounding objects. In contrast, our dataset introduces audio-visual camouflaged scenes with referring and reasoning expressions, requiring models to jointly understand sound, vision, and language.

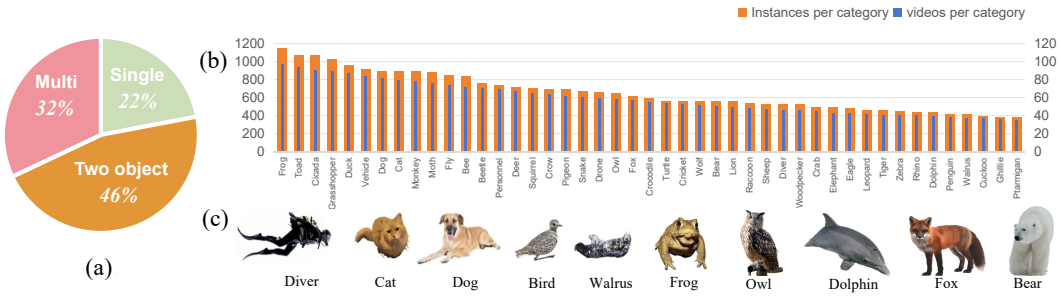

Figure 3: Dataset statistics and examples of our proposed **R2-AVCOSBench**. (a) The distribution of videos with different numbers of annotated objects. (b) The distribution of annotated instances and video clips across diverse object categories, showing a natural long-tailed characteristic. (c) Representative examples of object categories.

Reasoning-based Perception. With the rapid development of MLLMs, visual perception has shifted from simple semantic recognition to instruction following and reasoning-based understanding. LISA [17] introduces a `[SEG]` token to connect MLLMs with a mask decoder [18], enabling reasoning-based segmentation. VideoLISA [19] and VISA [20] further introduce temporal modeling and extend this paradigm to video reasoning segmentation. Although these methods show strong reasoning ability, they mainly focus on vision-language or video-language settings, without modeling audio content for segmentation-target reasoning, and are usually designed for natural images rather than camouflaged scenes. AnyRef [21] uses audio and images as referring inputs for segmentation, but it is mainly designed for static images and struggles with audio reasoning in dynamic videos. To this end, we propose referring and reasoning-based audio-visual camouflaged object segmentation.

3 Benchmark: R2-AVCOSBench

3.1 Audio-Visual Camouflaged Video Collection

The video sources consist of three main parts: i) real-world wildlife videos under the *Creative Commons License*, where we use techniques from [6, 7, 22] to align audio and visual snippets with intended semantics; ii) open-source audio-visual datasets containing diverse sounding animals and dynamic natural scenes; and iii) additional web videos of camouflaged animals with reliable audio-visual correspondence. We select videos from the above sources according to two main standards. First, we encourage to select videos that contain audio content with informative or reasoning-based meanings and corresponding identifiable camouflaged objects in the scene. Second, they should involve complex scenes with multiple objects, enabling diverse referring and reasoning expressions, e.g., referring to or reasoning about multiple camouflaged targets. As shown in Fig. 3(a), 78% of videos contain two or more objects, and 32% contain three or more objects. To ensure the diversity of referred and reasoned objects, we select 45 categories of camouflaged targets, including naturally camouflaged animals and a small number of human-related or artificial objects. Specifically, we include 16 terrestrial animals, 15 flying animals, 5 amphibians and reptiles, 4 aquatic animals, and 5 human-related or artificial categories, including vehicle, personnel, drone, diver, and ghillie. Fig. 3 provides detailed dataset statistics and representative examples. After reviewing about 10,871 candidates, we selected 2,654 videos that meet the above standards to construct R2-AVCOSBench, covering diverse real-world audio-visual camouflaged scenarios. We then annotate candidate camouflaged objects with expressions and bounding boxes at their first appearance, which are further used for video mask annotation of the camouflaged objects referred to or reasoned about by the expressions. Next, we introduce the annotation process.

3.2 Referring and Reasoning Expression and Mask Annotation

The expression annotation contains two types: *referring expressions* and *reasoning expressions*. Referring expressions identify the target camouflaged object using explicit audio-visual or temporal cues, while reasoning expressions require inferring the target from sound semantics, visual camouflage,

Table 1: Statistical comparison between the newly proposed R2-AVCOSBench and representative COS datasets. In **Content**, Text denotes textual guidance, including attribute text, category names, or natural-language expressions. In **Capability**, Ref., Reason., and Expl. denote whether the dataset supports referring expressions, reasoning expressions, and explanations, respectively. In **Statistics**, Text, Ref., Reason., Expl., and Cat. denote the numbers of textual guidance/prompts, referring expressions, reasoning expressions, explanations, and categories, respectively.

Dataset	Venue	Content				Capability			Statistics						
		Image	Video	Audio	Text	Ref.	Reason.	Expl.	Img./Vid.	Frame	Text	Ref.	Reason.	Expl.	Cat.
CAMO [23]	[CVIU'19]	✓	✗	✗	✗	✗	✗	✗	1,250	✗	✗	✗	✗	✗	8
COD10K [2]	[CVPR'20]	✓	✗	✗	✗	✗	✗	✗	5,066	✗	✗	✗	✗	✗	78
CHAMELEON [24]	[2018]	✓	✗	✗	✗	✗	✗	✗	76	✗	✗	✗	✗	✗	–
NC4K [9]	[CVPR'21]	✓	✗	✗	✗	✗	✗	✗	4,121	✗	✗	✗	✗	✗	–
COD-TAX [12]	[ECCV2024]	✓	✗	✗	✓	✗	✗	✗	4,040	✗	4,040	✗	✗	✗	–
OVCamo [13]	[ECCV2024]	✓	✗	✗	✓	✗	✗	✗	11,483	✗	11,483	✗	✗	✗	–
Text-COD [4]	[PR'26]	✓	✗	✗	✓	✗	✗	✗	10513	✗	10513	✗	✗	✗	–
CAD2016 [10]	[ECCV'16]	✗	✓	✗	✗	✗	✗	✗	9	400	✗	✗	✗	✗	–
MoCA-Mask [11]	[CVPR'22]	✗	✓	✗	✗	✗	✗	✗	141	37,241	✗	✗	✗	✗	67
R2-AVCOSBench	[Ours]	✗	✓	✓	✓	✓	✓	✓	**2,654**	**21,232**	**32,329**	**17,543**	**14,786**	**14,786**	**45**

scene context, and common-sense knowledge. Each expression is written in natural language and linked to one or more pixel-level masks. More word-level analysis is provided in Appendix B.

I) Referring Expressions. Referring expressions are designed to directly indicate the target camouflaged object in the audio-visual scene. The annotation process follows these rules: 1) Expressions should rely on audio-visual cues rather than pure visual appearance alone, since camouflaged targets are often difficult to distinguish from the background. 2) Expressions should integrate informative cues from audio, vision, and time. The audio dimension includes sound type, intensity, rhythm, and temporal order; the visual dimension includes appearance, location, motion, and spatial relations; the temporal dimension includes event order, such as "the one that sounds first" or "the object appearing later". 3) Expressions should be accurate and unambiguous. Specifically, the referred target should be uniquely determined by the expression, unless the expression explicitly refers to multiple camouflaged targets. 4) Each descriptive word should contribute to narrowing down the target scope, avoiding unnecessary or repetitive descriptions. 5) Subjective expressions, such as "the louder one" or "the more active one", are used only when the distinction is sufficiently clear to avoid ambiguity.

II) Reasoning Expressions. Reasoning expressions are designed to evaluate whether models can go beyond direct referring and infer camouflaged targets from higher-level audio-visual understanding. Unlike referring expressions, they usually do not directly provide the target category, appearance, or location, but describe the target via indirect semantic cues, behaviors, sound meanings, or scene-level implications. The annotation follows these rules: 1) Expressions should emphasize sound content and meaning rather than merely sound occurrence. For example, instead of "the animal making a sound", we use "the hidden animal warning others", which requires understanding the sound semantics. 2) Expressions should encourage reasoning from audio-visual evidence, such as sound semantics, motion patterns, temporal events, or scene interactions. For example, "the concealed animal most likely calling for a mate" requires understanding the biological meaning of the sound, rather than simply localizing the sounding region. 3) Expressions should require camouflaged object understanding, connecting reasoning results with subtle visual evidence such as weak boundaries, background similarity, or partially exposed body parts. For example, "the hidden target likely resting in its natural habitat" requires reasoning over both scene context and camouflage cues. 4) Expressions should be logically grounded in the video, with a clear and verifiable reasoning chain from expression to target. 5) For complex reasoning expressions, we provide explanations describing why the camouflaged object satisfies the expression. These explanations help evaluate not only segmentation accuracy but also whether the model understands the underlying reasoning process.

Mask Annotation. After collecting referring and reasoning expressions, the annotation team proceeds to segment the camouflaged objects indicated by each expression in every video. Using the bounding box of each target object's first appearance, as a reference, annotators are required to carefully confirm the corresponding camouflaged object in the video, and then track and annotate its mask across all frames. For expressions referring to multiple camouflaged targets, annotators provide masks for all matched objects. For no-target expressions, no mask is assigned. Annotators preview the video, confirm the target objects, and sequentially track and segment the camouflaged targets. All masks are further checked and refined manually to ensure accurate boundaries for camouflaged objects.

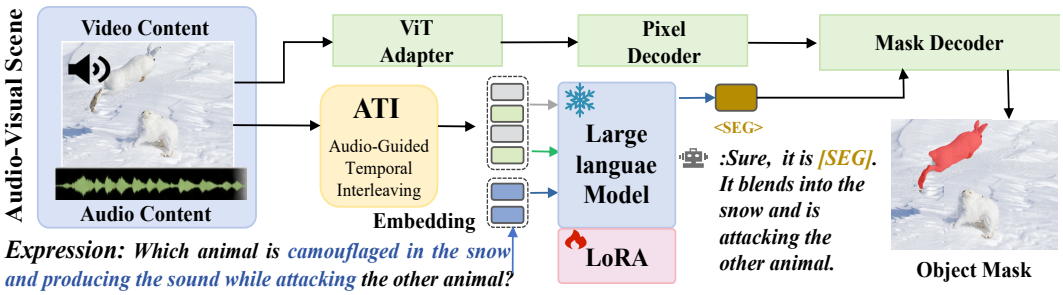

Figure 4: Overview of the proposed **CISA**. Given an audio-visual camouflaged video and a textual expression, CISA aligns audio and visual tokens via audio-guided temporal interleaving, generates a `[SEG]` token with the MLLM, and decodes it into the target mask with a reasoning explanation.

3.3 Dataset Statistics and Analysis

Table 1 compares R2-AVCOSBench with representative COS datasets. Existing image-level datasets, such as CAMO [23], COD10K [2], CHAMELEON [24], and NC4K [9], provide pixel-level annotations for static camouflaged objects, while video-level datasets such as CAD2016 [10] and MoCA-Mask [11] further introduce temporal information. Recent text-guided COD datasets, including COD-TAX [12], OVCamo [13], and Text-COD [4], incorporate textual guidance, but remain limited to image-level visual perception without audio cues, video dynamics, reasoning expressions, or explanatory annotations. In contrast, R2-AVCOSBench is the first benchmark to unify audio-visual camouflaged videos, text-based referring and reasoning expressions, pixel-level masks, and reasoning explanations. As summarized in Table 2, R2-AVCOSBench contains 2,654 audio-visual camouflaged videos, 21,232 annotated frames, and 32,329 expressions, including 17,543 referring expressions and 14,786 reasoning expressions. It covers 45 categories, 5.9 hours of videos, and a 6:4 train/test split with 19,397/12,932 expressions. These statistics show that R2-AVCOSBench extends COS from visual-only perception to audio-visual, text-guided referring and reasoning segmentation.

Table 2: More dataset statistics of R2-AVCOSBench.

Statistic	Number
Words per Expression	10.4
Frames per Video (mean)	8.0
Expression Types	2
Expression Forms	Ref. / Reason.
Frames (min/max)	4/28
Obj. per Expr. (mean)	1.05
Duration	5.9 h
Videos (train/test)	1,592/1,062
Frames (train/test)	12,739/8,493
Expr. (train/test)	19,397/12,932

4 Baseline Model: CISA

To support the proposed referring and reasoning audio-visual camouflaged object segmentation task, we propose **Camouflaged Instructed Segmentation Assistant (CISA)** as a baseline framework. As shown in Fig. 4, CISA consists of two main components: a Multimodal Large Language Model (MLLM) for audio-visual-text understanding and response generation, and a mask head for camouflaged object segmentation. The MLLM includes an audio encoder, a vision encoder, and a Large Language Model (LLM), while the mask head consists of a ViT-Adapter [25], a pixel decoder, and a mask decoder [26].

Given an audio-visual video, CISA encodes video frames and audio content into corresponding visual tokens and audio tokens. To achieve fine-grained audio-visual synchronization, we design Audio-Guided Temporal Interleaving (ATI), which divides audio tokens into temporally aligned clips according to sampled video frames and interleaves them with visual tokens. The resulting sequence forms the Audio-Visual Content Tokens. This strategy establishes temporal correspondence between audio cues and visual frames without introducing additional trainable fusion parameters, which is particularly useful for camouflaged scenes where visual evidence is weak and targets are highly similar to the background. When multiple visually hidden candidates appear in a scene, accurate audio-visual synchronization helps determine which camouflaged object is consistent with both the sound cue and the reasoning expression. For referring and reasoning expressions, we employ the text embedding module to encode textual instructions, producing the Referring-and-Reasoning Expression Tokens. The Audio-Visual Content Tokens and Referring-and-Reasoning Expression Tokens are then jointly fed into the MLLM. The MLLM generates textual responses and produces a `[SEG]` token representing the target camouflaged object. This `[SEG]` token is further passed to the mask decoder together with dense visual features from the mask head to predict the target mask.

Audio-Guided Temporal Interleaving. In audio-visual camouflaged videos, targets are often highly similar to the surrounding background, making them difficult to localize from single-frame appearance alone. Audio cues can provide critical semantic and temporal hints, but their effectiveness depends on precise temporal correspondence with video frames. To this end, we design ATI to construct frame-level aligned audio-visual content representations. For a video of arbitrary length, we uniformly sample N frames from the beginning to the end. Each sampled frame is processed by the vision encoder to obtain L_v visual tokens, denoted as $V = \{v_1, v_2, \ldots, v_N\}$, where $v_i \in \mathbb{R}^{L_v \times d}$ represents the visual tokens of the i-th sampled frame and d denotes the token dimension. For the audio stream, we use the audio encoder to extract audio tokens $A \in \mathbb{R}^{L_A \times d}$, where L_A is the total number of audio tokens. To establish temporal correspondence between audio and sampled frames, we partition the audio tokens according to the temporal intervals of the sampled frames, obtaining $\{a_1, a_2, \ldots, a_N\}$, where a_i denotes the audio token clip temporally aligned with the i-th sampled frame. We then interleave each visual token sequence with its corresponding audio clip to form the audio-visual content tokens $C_{\text{av}} = \{v_1, a_1, v_2, a_2, \ldots, v_N, a_N\}$.

In comparison, VideoLLaMA [27, 28] directly concatenates audio and visual tokens sequentially, $\{v_1, \ldots, v_N, a_1, \ldots, a_N\}$, which makes frame-level alignment difficult when the video FPS varies. Besides, video-SALMONN [29] concatenates audio and visual tokens along the feature dimension and maps them with a weight matrix, introducing fusion-specific parameters and potentially disturbing the pretrained MLLM feature space. Moreover, AVSBench [6] adopts a sliding-window strategy to capture local temporal context but still lacks explicit frame-level target correspondence.

Training Objectives. Following previous work [17], we employ cross-entropy loss for text generation, DICE loss, and binary cross-entropy loss for segmentation. The final training objective is:

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}. \tag{1}$$

where λ_{text}, λ_{bce}, and λ_{dice} are set to 1.0, 2.0, and 0.5, respectively, following LISA [17].

5 Experiment

Dataset. As described in Section 3, we use the R2-AVCOS benchmark with its train/test splits. Moreover, following previous works [17, 30, 20, 19], we additionally leverage training data, including: 1) semantic segmentation datasets: ADE20K [31], COCO-Stuff [32], PASCAL-Part [33], PACO-LVIS [34]; 2) referring segmentation datasets: RefCOCO, RefCOCO+ [35], RefCOCOg [36], ReasonSeg [17]; 3) referring video segmentation datasets: Refer-YouTube-VOS [37], Refer-DAVIS-17 [38], MeViS [39], ReVOS [20]; 4) audio-visual segmentation datasets: Ref-AVS Bench [7].

Implementation Details. We adopt InternVL2-1B [40] as the Multimodal Large Language Model (MLLM), where InternViT-300M-448px [40] serves as the vision encoder and Qwen2-0.5B-Instruct [41] serves as the Large Language Model (LLM). Whisper-large-v3 [42] is used as the audio encoder, followed by an audio MLP to project audio features into the LLM feature space. For the mask head, we use ViT-Adapter [25], a pixel decoder, and a mask decoder following Mask2Former [26]. To reduce memory consumption, we remove the Injectors in ViT-Adapter and remove the self-attention module in each mask-decoder transformer block, since only one target query is used. We denote this model as CISA. Following previous works [43, 44], since InternVL2 does not natively support audio input, we first conduct an audio-text alignment stage. Specifically, we use Automatic Speech Recognition datasets [43] and audio caption datasets [44] to align audio features with the LLM feature space. During this stage, only the audio MLP is trained, while all other parameters are frozen. In the subsequent CISA training stage, we use LoRA [45] to fine-tune the LLM, train all parameters of the mask head, and freeze the remaining parameters. We train with AdamW (learning rate 1×10^{-4}, batch size 8) on four NVIDIA RTX 4090 GPUs.

Evaluation Metrics. Following prior video segmentation benchmarks [7, 46, 47, 39, 48], we adopt $\mathcal{J}\&\mathcal{F}$ for evaluating camouflaged mask quality, where $\mathcal{J}$ measures region similarity and $\mathcal{F}$ evaluates contour accuracy. For no-target expressions, the score is set to 1 if the prediction is empty and 0 otherwise [49]. Following previous explanation evaluation practice [50], we use METEOR to assess generated explanations for reasoning expressions. More details on the evaluation metrics and human evaluation of explanations are provided in the Appendix A.

5.1 R2-AVCOSBench Benchmark Results

Table 4: Ablation studies on **audio-visual temporal modeling** and **mask head design**. ATI denotes Audio-Guided Temporal Interleaving; ViT-A denotes ViT-Adapter; Self-Attent. denotes self-attention.

(a) Audio-visual temporal modeling

ID	Fusion Strategy	Referring	Reasoning	All
I	Cross-Attention	43.2	37.6	40.5
II	Sequential Concat [27, 28]	45.1	39.4	42.3
III	Concat + MLP [29]	45.7	40.1	42.9
IV	Sliding Window [6]	46.4	41.2	43.8
V	**ATI**	**48.3**	**42.8**	**45.8**

(b) Mask head design (All Expressions Setting)

ID	Mask Head	ViT-A	Injectors	Self-Attent.	$\mathcal{J}\&\mathcal{F}$	FPS
I	SAM [18]	✗	✗	✗	45.1	3.6
II	Mask2Former	✗	✗	✓	43.7	18.9
III	Mask2Former	✓	✓	✓	46.0	12.4
IV	Mask2Former	✓	✗	✓	45.8	16.8
V	**Mask2Former**	✓	✗	✗	45.8	**22.7**

Table 3: Benchmark results on **R2-AVCOSBench**.

Method	All Expressions			Referring			Reasoning			METEOR
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	
AVSBench [6]	25.2	23.8	26.6	27.1	25.5	28.7	22.9	21.8	24.0	-
AVSegFormer [51]	28.4	26.9	29.9	30.6	28.9	32.3	25.8	24.6	27.0	-
LMPM [39]	29.1	27.6	30.6	31.4	29.7	33.1	26.4	25.1	27.7	-
MUTR [52]	35.4	33.8	37.0	38.0	36.2	39.8	32.3	30.9	33.7	-
EEMC [7]	33.1	31.5	34.7	36.2	34.4	38.0	29.4	28.0	30.8	-
LISA-7B [17]	36.7	35.0	38.4	39.5	37.6	41.4	33.4	31.9	34.9	22.8
LISA-13B [17]	39.2	37.5	40.9	42.1	40.2	44.0	35.8	34.2	37.4	27.2
OISA [8]	40.3	38.5	42.1	43.2	41.3	45.1	36.8	35.2	38.4	28.3
VideoLISA [19]	41.0	39.2	42.8	44.0	42.0	46.0	37.4	35.8	39.0	28.9
VISA [20]	42.1	40.3	43.9	45.0	43.0	47.0	38.6	37.0	40.2	29.8
CISA-1B (ours)	**45.8**	**44.0**	**47.5**	**48.3**	**46.6**	**50.0**	**42.8**	**41.0**	**44.6**	**32.4**

Quantitative results. Table 3 reports the benchmark results on our dataset. We compare CISA with representative methods from audio-visual segmentation, referring video segmentation, referring audio-visual segmentation, and large multimodal model based reasoning segmentation. Since most methods do not natively support our task, we adapt them under a unified protocol: audio-visual segmentation methods [6, 51] use expressions as target queries; referring video segmentation methods [39, 52] concatenate audio features with their original language-video queries; referring audio-visual segmentation methods [7, 8] are retrained on R2-AVCOSBench; and large multimodal model based reasoning segmentation methods [17, 19, 20] use the same audio-text alignment module as CISA while keeping their original segmentation strategies. Methods without explanation generation are evaluated only on mask quality, with METEOR marked as "-". As shown in Table 3, CISA achieves the best performance across all expression types. The gains on referring and reasoning expressions indicate stronger audio-visual grounding and better integration of sound semantics, visual camouflage cues, and textual intent, while the higher METEOR score suggests more faithful explanations.

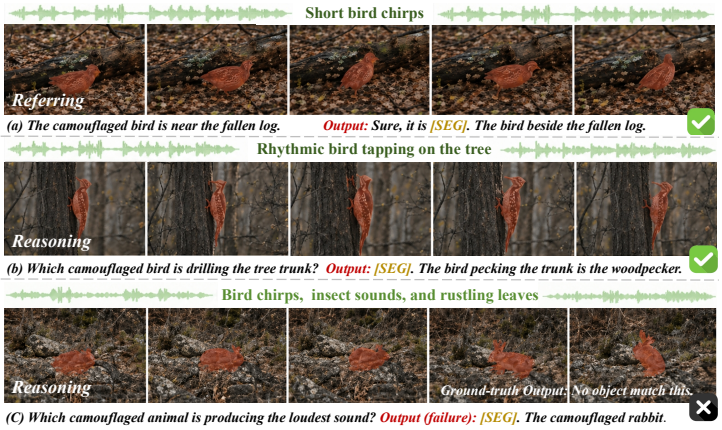

Figure 5: Qualitative case study on R2-AVCOSBench.

Qualitative Results and Failure Cases. Fig. 5 presents several success and failure cases of CISA. Case (a) shows effective alignment between bird chirps and visual content, leading to accurate segmentation of the camouflaged bird specified by the referring expression. Case (b) further illustrates reasoning ability by associating rhythmic tapping with the camouflaged bird pecking the tree trunk. In contrast, Case (c) reveals a no-match failure under complex audio-visual distractors. Although the audio contains bird chirps, insect sounds, and rustling leaves, the model still segments the silent camouflaged rabbit. This highlights the difficulty of rejecting unmatched queries and motivates future work on sound-relevant target discrimination and intention-aware reasoning.

5.2 Ablation Study

Ablation Study on Audio-Visual Temporal Modeling. As shown in Table 4(a), cross-attention performs worst, indicating that implicit token interaction is insufficient for fine-grained audio-visual

Table 6: Comparison on external referring and reasoning segmentation benchmarks. Left: performance on the testing set of **Ref-AVS** [7]. Right: cIoU results on **refCOCO** [35, 36] and **ReasonSeg** [17].

(a) Ref-AVS

Method	Seen		Unseen		Mix		Null
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{S}$ (↓)
AVSBench [6]	23.2	51.1	32.4	54.7	27.8	52.9	20.8
AVSegFormer [51]	33.5	47.0	36.1	50.1	34.8	48.6	17.1
GAVS [53]	28.9	49.8	29.8	49.7	29.4	49.8	19.0
ReferFormer [54]	31.3	50.1	30.4	48.8	30.9	49.5	17.6
R2VOS [55]	25.0	41.0	27.9	49.8	26.5	45.4	18.3
EEMC [7]	34.2	51.3	49.5	64.8	41.9	58.1	**7.0**
CISA-1B (ours)	**44.7**	**54.6**	**49.8**	**65.1**	**46.5**	**59.4**	9.8

(b) Referring and Reasoning Segmentation

Method	refCOCO			refCOCO+			refCOCOg		ReasonSeg	
	val	testA	testB	val	testA	testB	val_U	test_U	val	test
LAVT [56]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	–	–
ReLA [49]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	–	–
LISA-7B [17]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	54.0	48.4
LISA-13B [17]	–	–	–	–	–	–	–	–	62.9	51.1
VISA-7B [20]	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4	57.8	–
VideoLISA-3.8B [19]	73.8	76.6	68.8	63.4	68.8	56.2	68.3	68.8	**67.1**	**54.4**
CISA-1B (ours)	**74.2**	**77.3**	**69.6**	65.8	**71.4**	**58.0**	**69.6**	**70.3**	59.2	50.1

alignment. Sequential concatenation [27, 28] and MLP-based fusion [29] improve performance, while sliding-window alignment further benefits temporal correspondence. Our ATI achieves the best results across all splits, validating temporal interleaving for audio-visual grounding and reasoning.

Ablation Study on Mask Head Design. Table 4(b) compares different mask heads. SAM provides competitive masks but is slow due to its heavy backbone. Mask2Former improves efficiency, while the plain head lowers mask quality. ViT-Adapter improves representation, and Injectors with self-attention further enhance feature interaction. Removing them increases speed with slight accuracy loss. We therefore adopt the final Mask2Former-based design.

Table 5: Modality ablation analysis. $\mathcal{J}\&\mathcal{F}$ is the default metric.

Setting	Vision	Text	Audio	Referring	Reasoning	All
Vision	✓	✗	✗	42.1	37.6	39.4
w/o Text	✓	✗	✓	44.8	40.3	41.1
Full (ours)	✓	✓	✓	**48.3**	**42.8**	**45.8**

Ablation Study on Inputs. Table 5 studies audio and text contributions with vision always available. Removing audio degrades performance, especially on reasoning expressions, showing the value of sound semantics. Removing text causes a larger drop, since expressions define user intention. The full model performs best, indicating that R2-AVCOS requires joint visual, audio, and textual cues.

5.3 Results on Related Tasks

Results on Ref-AVS. As shown in Table 6(a), CISA achieves strong performance on Ref-AVS [7], outperforming EEMC [7] on the Seen, Unseen, and Mix splits. This indicates that CISA transfers well to Ref-AVS and generalizes to unseen categories. Although its score on the Null split is not the best, the gains on standard referring splits demonstrate robust audio-visual understanding.

Table 7: Results on **referring/reasoning video segmentation** [39, 37, 38, 20]. $\mathcal{J}\&\mathcal{F}$ is the default metric, and $\mathcal{R}$ denotes the robustness score.

Method	MeViS	R-YTVOS	R-DAVIS_{17}	ReVOS	$\text{ReVOS}_{(\mathcal{R})}$
ReferFormer [54]	31.0	62.9	61.1	14.9	4.9
MTTR [57]	30.0	55.3	–	25.5	5.6
LMPM [39]	37.2	–	–	26.4	3.2
LISA-7B [17]	–	50.2	58.4	40.9	9.3
TrackGPT-7B [58]	40.1	56.4	63.2	43.6	11.6
VideoLISA-3.8B [19]	42.3	61.7	67.7	–	–
VISA-7B [20]	44.5	**63.0**	**70.4**	47.1	15.3
CISA-1B (ours)	**45.2**	61.8	67.2	**47.4**	**18.7**

Results on Referring Image Segmentation. Table 6(b) reports results on the RefCOCO series [35, 36] and ReasonSeg [17]. Our model achieves competitive performance, showing that it preserves strong referring segmentation ability despite being designed for audio-visual camouflaged scenes. On ReasonSeg, it lags behind larger reasoning segmentation models, likely due to its smaller language model and task focus on audio-visual camouflage. Even so, it maintains reasonable reasoning ability with a compact model size.

Results on Referring Video Segmentation. As shown in Table 7, CISA obtains competitive results on MeViS [39], R-YTVOS [37], and R-DAVIS_{17} [38]. Although some video-specific models perform better on individual datasets, our model remains comparable while additionally supporting audio-visual reasoning and camouflaged object segmentation. On ReVOS [20], it performs strongly in both segmentation quality and robustness, suggesting good generalization to complex referring and reasoning video segmentation tasks. More COS-related results are provided in the Appendix D.

5.4 Additional Analyses

To complement the main experiments, we further analyze the proposed benchmark and baseline from three perspectives. First, controlled audio and query perturbations examine the sensitivity of CISA to temporal correspondence and consistency among modalities. Second, independent annotation assessments quantify agreement on target identity, reasoning validity, explanation consistency, and no-target correctness. Finally, we compare CISA with direct SAM3 prompting and an MLLM+SAM3 pipeline to assess performance against alternative segmentation approaches. For the pipeline comparison, we use the same reasoning backbone and audio encoder as CISA.

Cross-Modal Consistency. We perturb audio or textual queries while keeping the checkpoint and remaining inputs fixed. For audio mismatch, we select audio from another same-category video of closest duration. As shown in Table 8, reasoning $\mathcal{J}\&\mathcal{F}$ decreases from 42.8 to 39.0 under audio shuffling and to 38.9 under same-category audio mismatch. Replacing the query with a same-category expression from another video reduces the score to 39.3. For this query-sensitivity test, predictions are evaluated against the original target masks. These results support the importance of temporal audio-visual correspondence and query consistency.

Table 8: Cross-modal consistency on the reasoning subset of R2-AVCOSBench. Results are measured by $\mathcal{J}\&\mathcal{F}$.

Condition	Reasoning $\mathcal{J}\&\mathcal{F}$
Aligned audio and correct query	**42.8**
Without audio	40.3
Shuffled audio segments	39.0
Same-category mismatched audio	38.9
Same-category mismatched query	39.3

Annotation Reliability. Three independent annotators assess 300 reasoning cases for target identity and uniqueness, reasoning validity, explanation consistency, and no-target correctness. As shown in Table 9, agreement ranges from 89.7% to 96.0%, with Fleiss' κ ranging from 0.81 to 0.92. An additional annotator reviews disagreements, and unresolved cases are revised or removed.

Table 9: Annotation reliability on 300 reasoning cases.

Evaluation Aspect	Agreement (%)	Fleiss' κ
Target identity and uniqueness	94.7	0.89
Reasoning validity	91.3	0.84
Explanation consistency	89.7	0.81
No-target correctness	96.0	0.92

Comparison with SAM3. Table 10 reports results for direct text prompting and an MLLM+SAM3 pipeline that first resolves a target description and then performs segmentation. Direct prompting achieves 36.4 referring $\mathcal{J}\&\mathcal{F}$. Using the same InternVL2-1B backbone and Whisper-large-v3 audio encoder as CISA, the two-stage pipeline obtains 37.1 overall $\mathcal{J}\&\mathcal{F}$, compared with 45.8 for CISA.

Table 10: Comparison with SAM3 and an MLLM+SAM3 pipeline on R2-AVCOSBench. All scores are $\mathcal{J}\&\mathcal{F}$.

Method	Referring	Reasoning	Overall
SAM3 (direct)	36.4	–	–
MLLM+SAM3	39.2	34.8	37.1
CISA-1B (ours)	**48.3**	**42.8**	**45.8**

6 Conclusion and Discussion

We introduce R2-AVCOSBench, a large-scale benchmark for referring and reasoning audio-visual camouflaged object segmentation. Unlike previous COS datasets that mainly focus on visual-only image or video perception, R2-AVCOSBench introduces audio-visual camouflaged scenes with text-based referring and reasoning expressions, enabling deeper understanding of sound semantics, visual camouflage cues, scene context, and user intention. To address this task, we propose CISA, a simple yet effective MLLM-based segmentation baseline, enabling effective audio-visual grounding and intention-aware camouflaged object segmentation. Extensive experiments demonstrate its effectiveness on our dataset and related referring and reasoning segmentation benchmarks.

Limitations and Future Directions. Although CISA shows promising results, several challenges remain. In complex scenes with multiple sound sources and highly concealed objects, disentangling audio events and associating them with the correct camouflaged targets is still difficult. Future work may explore stronger audio-visual fusion, reasoning mechanisms, larger-scale pre-training, and interactive settings to better handle ambiguous, no-match, and multi-target cases.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants U25A20431, 62432003, and 92567302.

References

[1] Abduljalil Radman and Jorma Laaksonen. Tsam: Temporal sam augmented with multimodal prompts for referring audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23947–23956, June 2025.

[2] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2777–2787, June 2020.

[3] Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3597–3610, 2025.

[4] Zefeng Chen, Yunqi Xue, Zhijiang Li, Philip Torr, and Jindong Gu. Text-guided camouflaged object detection. *Pattern Recognition*, 170:112058, 2026.

[5] Yuli Zhou, Guolei Sun, Yawei Li, Guo-Sen Xie, Luca Benini, and Ender Konukoglu. When sam2 meets video camouflaged object segmentation: A comprehensive evaluation and adaptation. *Visual Intelligence*, 3:10, 2025.

[6] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-Visual Segmentation. In *Eur. Conf. Comput. Vis*, 2022.

[7] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes. In *Eur. Conf. Comput. Vis*, 2024.

[8] Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22575–22585, 2025.

[9] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11591–11601, 2021.

[10] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *Computer Vision – ECCV 2016*, pages 433–449. Springer International Publishing, 2016.

[11] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13864–13873, June 2022.

[12] Hong Zhang, Yixuan Lyu, Qian Yu, Hanyang Liu, Huimin Ma, Ding Yuan, and Yifan Yang. Unlocking attributes' contribution to successful camouflage: A combined textual and visual analysis strategy. In *Computer Vision – ECCV 2024*, pages 315–331, Cham, 2025. Springer Nature Switzerland.

[13] Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. In *Computer Vision – ECCV 2024*, pages 476–495, 2024.

[14] Ruohao Guo, Xianghua Ying, Yaru Chen, Dantong Niu, Guangyao Li, Liao Qu, Yanyu Qi, Jinxing Zhou, Bowei Xing, Wenzhen Yue, Ji Shi, Qixun Wang, Peiliang Zhang, and Buwen Liang. Audio-visual instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13550–13560, June 2025.

[15] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2022.

[16] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. AVQA: A Dataset for Audio-Visual Question Answering on Videos. In *ACM Int. Conf. Multimedia*, 2022.

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2024.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Int. Conf. Comput. Vis*, 2023.

[19] Zechen Bai, Tong He, Haiyang Mei, Pichao WANG, Ziteng Gao, Joya Chen, liulei, Zheng Zhang, and Mike Zheng Shou. One Token to Seg Them All: Language Instructed Reasoning Segmentation in Videos. In *Adv. Neural Inform. Process. Syst*, 2024.

[20] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning Video Object Segmentation via Large Language Models. In *Eur. Conf. Comput. Vis*, 2024.

[21] Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal Instruction Tuned LLMs with Fine-grained Visual Perception. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2024.

[22] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A Large-scale Audio-Visual Dataset. In *IEEE Int. Conf. Acoust. Speech Signal Process*, 2020.

[23] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019.

[24] Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: CHAMELEON database. *Unpublished Manuscript*, 2(6):7, 2018.

[25] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *Int. Conf. Learn. Represent*, 2023.

[26] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2022.

[27] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proc. of the Conf. on Empirical Methods in Nat. Lang. Process*, 2023.

[28] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

[29] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models. In *Int. Conf. Mach. Learn*, 2024.

[30] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. ViLLa: Video Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2407.14500*, 2024.

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2017.

[32] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and Stuff Classes in Context. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2018.

[33] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2014.

[34] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. PACO: Parts and Attributes of Common Objects. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2023.

[35] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling Context in Referring Expressions. In *Eur. Conf. Comput. Vis*, 2016.

[36] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2016.

[37] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *Eur. Conf. Comput. Vis*, 2020.

[38] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video Object Segmentation with Language Referring Expressions. In *ACCV*, 2019.

[39] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *Int. Conf. Comput. Vis*, 2023.

[40] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*, 2024.

[41] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen Technical Report. *arXiv*, 2023.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Int. Conf. Mach. Learn*, 2023.

[43] Guoguo Chen, Shuzhou Chai, et al. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, 2021.

[44] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-ACD: A Large-scale Dataset for Audio-Language Representation Learning. In *ACM Int. Conf. Multimedia*, 2024.

[45] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Int. Conf. Learn. Represent*, 2022.

[46] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A New Dataset for Video Object Segmentation in Complex Scenes. In *Int. Conf. Comput. Vis*, 2023.

[47] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv*, 2025.

[48] Kaining Ying, Hengrui Hu, and Henghui Ding. MOVE: Motion-guided few-shot video object segmentation. In *Int. Conf. Comput. Vis*, 2025.

[49] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized Referring Expression Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2023.

[50] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Assoc. Comput. Linguist. Worksh*, 2005.

[51] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. AVSegFormer: Audio-Visual Segmentation with Transformer. In *AAAI*, 2024.

[52] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by Multi-Modality: A Unified Temporal Transformer for Video Object Segmentation. In *AAAI*, 2024.

[53] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting Segmentation with Sound Is Generalizable Audio-Visual Source Localizer. In *AAAI*, 2024.

[54] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as Queries for Referring Video Object Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2022.

[55] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust Referring Video Object Segmentation with Cyclic Structural Consensus. In *Int. Conf. Comput. Vis*, 2023.

[56] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2022.

[57] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-End Referring Video Object Segmentation with Multimodal Transformers. In *IEEE Conf. Comput. Vis. Pattern Recog*, 2022.

[58] Jiawen Zhu, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Huchuan Lu, Yifeng Geng, and Xuansong Xie. Tracking with Human-Intent Reasoning. *arXiv*, 2023.

[59] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2160–2170, 2022.

[60] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4722, 2022.

[61] Yixuan Lyu, Hong Zhang, Yan Li, Hanyang Liu, Yifan Yang, and Ding Yuan. Uedg: Uncertainty-edge dual guided camouflage object detection. *IEEE Transactions on Multimedia*, 26:4050–4060, 2023.

[62] Haozhe Xing, Shuyong Gao, Hao Tang, Tsui Qin Mok, Yanlan Kang, and Wenqiang Zhang. Tinycod: Tiny and effective model for camouflaged object detection. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[63] Yu Liu, Haihang Li, Juan Cheng, and Xun Chen. Mscaf-net: A general framework for camouflaged object detection via learning multi-scale context-aware features. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4934–4947, 2023.

[64] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, HuaiXin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5557–5566, 2023.

[65] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22046–22055, 2023.

[66] Ze Song, Xudong Kang, Xiaohui Wei, Haibo Liu, Renwei Dian, and Shutao Li. Fsnet: Focus scanning network for camouflaged object detection. *IEEE Transactions on Image Processing*, 32:2267–2278, 2023.

[67] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson W. H. Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 781–789, 2023.

[68] Xihang Hu, Xiaoli Zhang, Fasheng Wang, Jing Sun, and Fuming Sun. Efficient camouflaged object detection network based on global localization perception and local guidance refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5452–5465, 2024.

[69] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[70] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[71] Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. Exploring deeper! segment anything model with depth perception for camouflaged object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 4322–4330, 2024.

[72] Tingran Wang, Zaiyang Yu, Jianwei Fang, Jinlong Xie, Feng Yang, Huang Zhang, Liping Zhang, Minghua Du, Lusi Li, and Xin Ning. Multidimensional fusion of frequency and spatial domain information for enhanced camouflaged object detection. *Information Fusion*, 117:102871, 2025.

[73] Xiaoqin Zhang, Zhenni Yu, Li Zhao, Deng-Ping Fan, and Guobao Xiao. Comprompter: Reconceptualized segment anything model with multiprompt network for camouflaged object detection. *Science China Information Sciences*, 68(1):112104, 2025.

[74] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8779–8788, October 2019.

[75] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7479–7489, June 2019.

[76] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3907–3916, June 2019.

[77] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 263–273. Springer International Publishing, 2020.

[78] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022.

[79] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 142–152. Springer International Publishing, 2021.

[80] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7284–7293, October 2019.

[81] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7177–7188, October 2021.

[82] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[83] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604. IEEE, 2009.

[84] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 12 2015.

[85] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 698–704. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

A Evaluation Metrics

A.1 Automatic Evaluation Metrics

Following prior video segmentation benchmarks [7, 46, 47, 39, 48], we evaluate mask quality using region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$. For a predicted mask P and the ground-truth mask G, $\mathcal{J}$ is computed as the intersection-over-union between the two masks:

$$\mathcal{J}(P, G) = \frac{|P \cap G|}{|P \cup G|}. \tag{2}$$

This metric measures the region-level overlap between the predicted and ground-truth masks.

The contour accuracy $\mathcal{F}$ evaluates boundary alignment between the predicted and ground-truth masks. It is computed from contour precision and contour recall:

$$\mathcal{F} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

A higher $\mathcal{F}$ score indicates better boundary localization and contour consistency. The final mask-quality score is calculated as the average of $\mathcal{J}$ and $\mathcal{F}$:

$$\mathcal{J}\&\mathcal{F} = \frac{\mathcal{J} + \mathcal{F}}{2}. \tag{4}$$

For each expression, we compute these metrics over the annotated frames and average the results. For no-target expressions, where no object in the video satisfies the given expression, the ground-truth mask is empty. Following the evaluation protocol of generalized referring segmentation [49], the score is set to 1 if the predicted mask is empty and 0 otherwise. This design explicitly evaluates whether the model can reject invalid or unmatched expressions, rather than always producing a mask.

For reasoning expressions with explanation annotations, following previous works [7, 37] on explanation and text generation evaluation [50], we further use METEOR to evaluate the generated explanations. METEOR measures the similarity between the generated explanation and the reference explanation by considering unigram matching, stemming, synonym matching, and word-order penalty. In our benchmark, METEOR complements mask-based metrics by assessing whether the model can provide an explanation that is semantically consistent with the annotated reasoning process.

A.2 Human Evaluation for Explanation Quality

Although METEOR [50] is used as an automatic reference-based metric for explanation evaluation, it mainly measures lexical overlap and cannot fully reflect whether an explanation is faithful to the audio-visual evidence or aligned with the intended target. Therefore, we further conduct a human evaluation for reasoning explanations.

Specifically, we randomly sample 300 reasoning-expression cases from the test set. Each case is independently evaluated by three human evaluators from two dimensions: **Explanation Credibility**

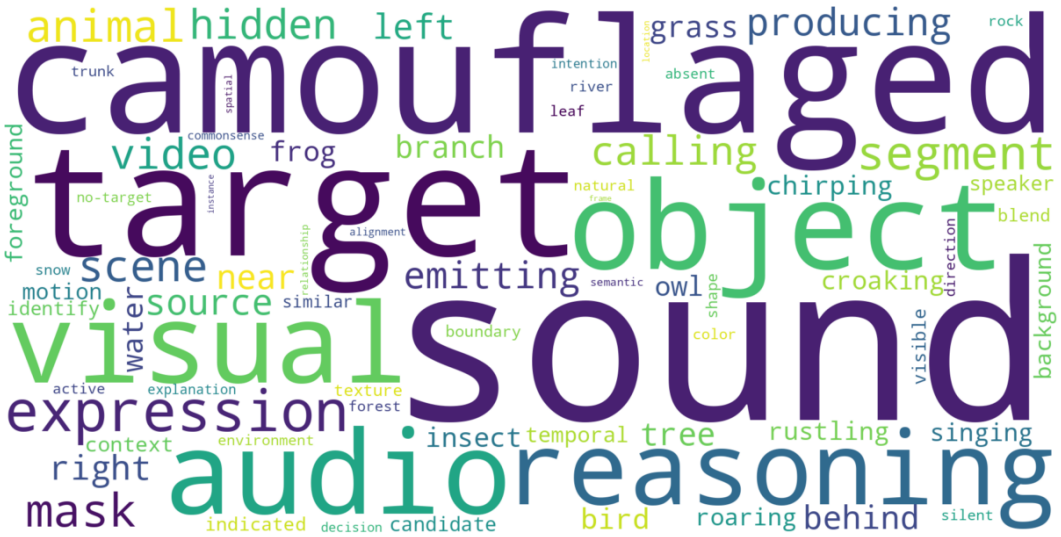

Figure 6: Word cloud of the top 100 words in R2-AVCOSBench.

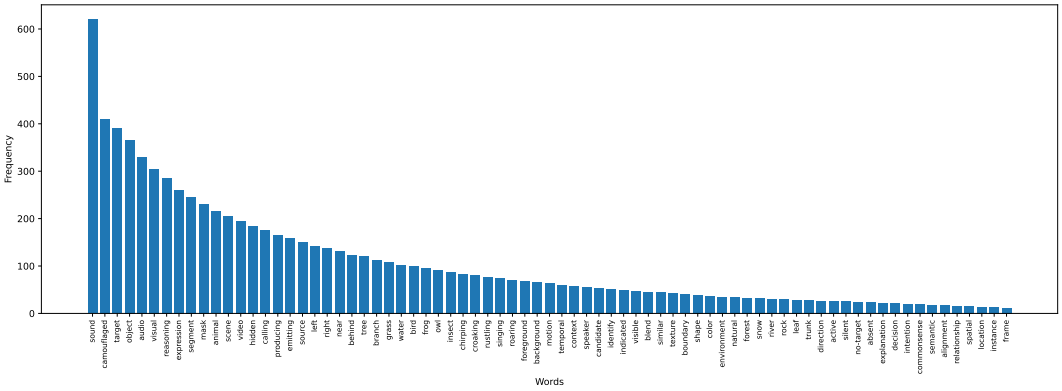

Figure 7: Word frequency of the top 75 words in R2-AVCOSBench.

(EC) and **Instruction Alignment (IA)**. EC measures whether the explanation is supported by the audio-visual content, scene context, object behavior, sound semantics, and camouflage-related visual evidence. IA measures whether the explanation correctly follows the user expression and is consistent with the predicted target mask. Each dimension is scored from 1 to 5, where 1 indicates poor quality and 5 indicates high quality. For each evaluator, the total score is computed as EC+IA, ranging from 2 to 10. The reported average score is computed over the three evaluators.

During evaluation, human evaluators are shown the video frames, audio description, user expression, predicted mask, generated explanation, and reference annotation. They are instructed to penalize explanations that hallucinate unsupported evidence, describe the wrong object, ignore the key reasoning intent, or are inconsistent with the predicted mask. This human evaluation complements METEOR by directly assessing whether the generated explanations are grounded in audio-visual evidence and aligned with the intended reasoning target.

Table 11: Human evaluation of generated explanations. Human1–Human3 denote three independent evaluators. Each score is reported as EC / IA / Total.

Method	Human1	Human2	Human3	Avg. ↑
LISA-7B [17]	2.6 / 2.9 / 5.5	2.8 / 3.1 / 5.9	2.7 / 3.0 / 5.7	5.7
LISA-13B [17]	3.1 / 3.4 / 6.5	3.3 / 3.5 / 6.8	3.2 / 3.6 / 6.8	6.7
VideoLISA [19]	3.4 / 3.6 / 7.0	3.6 / 3.5 / 7.1	3.5 / 3.7 / 7.2	7.1
VISA [20]	3.6 / 3.7 / 7.3	3.5 / 3.9 / 7.4	3.7 / 3.8 / 7.5	7.4
CISA-1B (ours)	**4.1 / 4.3 / 8.4**	**4.0 / 4.2 / 8.2**	**4.2 / 4.4 / 8.6**	**8.4**

B Word Analysis

We analyze the word distribution of R2-AVCOSBench using a word cloud and word-frequency statistics, as shown in Figures 6 and 7. The most frequent words include "sound", "camouflaged", "target", "object", "audio", "visual", and "reasoning", indicating that the benchmark is centered on audio-visual camouflaged target understanding. Words such as "calling", "emitting", "chirping", "croaking", "rustling", and "roaring" show that the expressions explicitly encode diverse sound semantics rather than merely indicating the presence of audio. Meanwhile, words related to visual and spatial grounding, such as "hidden", "scene", "near", "behind", "tree", "branch", and "background", suggest that the expressions require models to associate acoustic cues with camouflaged visual regions under complex scene contexts. In addition, terms such as "no-target", "absent", "explanation", "intention", and "commonsense" reflect the reasoning-oriented nature of R2-AVCOSBench, where models are expected to handle target absence and provide explanation-aware segmentation decisions. Overall, the word distribution demonstrates that R2-AVCOSBench contains rich and diverse expressions involving audio semantics, visual camouflage cues, spatial relations, temporal context, and commonsense reasoning.

C More Examples in R2-AVCOSBench

Figure 8 provides more examples from R2-AVCOSBench. The benchmark contains diverse audio-visual camouflaged scenes, where the target may be directly specified by a referring expression or implicitly inferred from sound semantics, spatial context, visual camouflage cues, and commonsense knowledge. These examples further illustrate the need for joint audio-visual grounding and reasoning-aware segmentation.

D Evaluation on Standard Image and Video COS Benchmarks

Motivation and protocol. Our main task, R2-AVCOS, differs fundamentally from conventional Camouflaged Object Segmentation. R2-AVCOS requires the model to segment camouflaged targets in audio-visual scenes according to user-provided referring or reasoning expressions, where the target may be specified by sound semantics, visual camouflage cues, temporal context, commonsense knowledge, or user intention. In contrast, conventional COS aims to segment camouflaged object(s) from images or videos without audio input, textual expressions, or reasoning-oriented instructions. Therefore, directly comparing R2-AVCOS performance on our benchmark with conventional COS results is not sufficient to evaluate the full task setting.

Nevertheless, standard COS benchmarks provide useful auxiliary evaluation for analyzing the generic camouflaged mask prediction ability of our model. To align our model with the conventional COS setting, we remove expression-specific and audio-specific requirements and use generic segmentation instructions. For image COS benchmarks, we provide the following prompt for each test image:

> USER: $\langle \mathrm{IMAGE} \rangle$ Please segment the camouflaged object in this image.
> ASSISTANT: Sure, it is $\langle \mathrm{SEG} \rangle$.

For video COS benchmarks, we use a similar video-level prompt:

> USER: $\langle \mathrm{VIDEO} \rangle$ Please segment the camouflaged object throughout the video.
> ASSISTANT: Sure, it is $\langle \mathrm{SEG} \rangle$.

When the evaluated benchmark does not provide audio, we use visual inputs only, or equivalently feed an empty audio placeholder when required by implementation. The predicted masks are compared with the corresponding ground-truth COS annotations. This protocol isolates the image-level and video-level camouflaged object segmentation ability of our model while keeping the evaluation goal consistent with prior COS methods.

Image COS datasets and metrics. For image-level evaluation, we follow the standard COS evaluation protocol on three widely used benchmarks: CAMO [23], COD10K [2], and NC4K [9]. For each dataset, we use the official test split and the provided pixel-level ground-truth masks. To be consistent with the COS literature, we report four commonly used image-level metrics, including



Figure 8: More examples in R2-AVCOSBench. The orange and red masks are only different visualization colors for better readability and do not indicate different annotation types or semantic meanings.

S-measure (S_m) [82], weighted F-measure (F_β^ω) [83], mean absolute error (MAE) [84], and E-measure (E_m) [85]. Higher values of S_m, F_β^ω, and E_m indicate better performance, while a lower value of MAE is better.

Video COS datasets and metrics. For video-level evaluation, we further evaluate our model on standard video camouflaged object segmentation benchmarks, such as CAD [10] and MoCA-Mask [11]. Different from image COS, video COS requires the model to produce temporally consistent frame-level masks across a video sequence. Following common VCOS protocols, We adopt seven evaluation metrics to measure the quality of predicted pixel-wise masks: S-measure (S_m), F-measure (F_β), weighted F-measure (F_β^ω), mean absolute error (MAE), E-measure (E_m), mean Dice (mDice), and mean IoU (mIoU).

Discussion. These COS-style evaluations should be viewed as auxiliary analyses rather than the main evidence for R2-AVCOS. Image COS evaluates whether the model can segment camouflaged objects from static visual appearance, while video COS further examines whether the model can maintain target consistency across frames. However, neither setting evaluates audio-content understanding,

Table 12: Quantitative comparison on standard image-level COS benchmarks. Higher values of S_m, $F_\beta^{\max}$, and $E_\phi^{\max}$ are better, while a lower value of M is better.

Method	Pub.	CAMO [23]				COD10K [2]				NC4K [9]			
		$S_m \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	MAE $\downarrow$
ZoomNet [59]	[CVPR'22]	0.820	0.805	0.889	0.066	0.837	0.777	0.896	0.029	0.852	0.826	0.903	0.044
SegMAR [60]	[CVPR'22]	0.816	0.803	0.884	0.071	0.833	0.774	0.906	0.034	0.841	0.826	0.907	0.046
UEDG [61]	[TMM'23]	0.863	0.856	0.929	0.048	0.858	0.812	0.934	0.025	0.879	0.864	0.935	0.035
TinyCOD [62]	[ICASSP'23]	0.822	0.807	0.899	0.066	0.811	0.742	0.903	0.036	0.843	0.817	0.910	0.047
MSCAF-Net [63]	[TCSVT'23]	0.873	0.867	0.937	0.046	0.865	0.823	0.936	0.024	0.887	0.874	0.942	0.032
FSPNet [64]	[CVPR'23]	0.856	0.846	0.928	0.050	0.851	0.794	0.930	0.026	0.879	0.859	0.937	0.035
FEDER [65]	[CVPR'23]	0.802	0.789	0.873	0.071	0.822	0.768	0.905	0.032	0.847	0.833	0.915	0.044
FSNet [66]	[TIP'23]	0.880	0.878	0.941	**0.041**	0.870	0.833	0.948	0.023	0.891	0.880	0.948	0.031
CRNet [67]	[AAAI'23]	0.735	0.707	0.830	0.092	0.733	0.636	0.845	0.049	–	–	–	–
PRNet [68]	[TCSVT'24]	0.872	0.867	0.930	0.050	0.873	0.839	0.943	0.022	0.891	0.881	0.942	0.031
CamoFormer [69]	[TPAMI'24]	0.817	0.801	0.883	0.068	0.838	0.786	0.928	0.029	0.854	0.829	0.908	0.042
ZoomNeXt [70]	[TPAMI'24]	0.874	0.873	0.931	0.047	0.887	0.856	0.945	**0.019**	0.892	0.884	0.941	**0.030**
DSAM [71]	[ACM MM'24]	0.832	0.834	0.920	0.061	0.845	0.805	0.930	0.033	0.872	0.864	0.942	0.040
MAMIFNet [72]	[IF'25]	0.872	0.870	0.935	0.045	0.869	0.826	0.940	0.023	0.890	0.878	0.943	0.031
COMPrompter [73]	[SCIS'25]	0.853	0.856	0.931	0.054	0.860	0.826	0.946	0.027	0.880	0.876	0.946	0.036
CISA (ours)	[Ours]	**0.881**	**0.887**	**0.943**	0.043	**0.905**	**0.875**	**0.954**	0.020	**0.908**	**0.898**	**0.950**	**0.029**

Table 13: Comparison with existing methods on video camouflaged object segmentation benchmarks. The best results are shown in **bold**.

Method	CAD [10]							MoCA-Mask [11]						
	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
EGNet[19] [74]	0.619	0.298	0.044	0.350	0.666	0.324	0.243	0.547	0.110	0.035	0.136	0.574	0.143	0.096
BASNet[19] [75]	0.639	0.349	0.054	0.394	0.773	0.393	0.293	0.561	0.154	0.042	0.173	0.598	0.190	0.137
CPD[19] [76]	0.622	0.289	0.049	0.357	0.667	0.330	0.239	0.561	0.121	0.041	0.152	0.613	0.162	0.113
PraNet[20] [77]	0.629	0.352	0.042	0.397	0.763	0.378	0.290	0.614	0.266	0.030	0.296	0.674	0.311	0.234
SINet[20] [2]	0.636	0.346	0.041	0.395	0.775	0.381	0.283	0.598	0.231	0.028	0.256	0.699	0.277	0.202
SINet-V2[21] [78]	0.653	0.382	0.039	0.432	0.762	0.413	0.318	0.588	0.204	0.031	0.229	0.642	0.245	0.180
PNS-Net[21] [79]	0.655	0.325	0.048	0.417	0.673	0.384	0.290	0.526	0.059	0.035	0.084	0.530	0.084	0.054
RCRNet[19] [80]	0.627	0.287	0.048	0.328	0.666	0.309	0.229	0.555	0.138	0.033	0.159	0.527	0.171	0.116
MG[21] [81]	0.594	0.336	0.059	0.375	0.692	0.368	0.268	0.530	0.168	0.067	0.195	0.561	0.181	0.127
STL-Net-LT[22] [11]	0.696	0.481	0.030	0.524	0.845	0.493	0.402	0.631	0.311	0.027	0.331	**0.759**	0.360	0.272
STL-Net-ST[22] [11]	0.696	0.471	0.031	0.515	0.827	0.484	0.392	0.637	0.304	0.027	0.328	0.734	0.356	0.271
ZoomNeXt[24] [70]	0.721	0.525	0.024	0.567	0.759	0.523	0.436	0.690	0.395	0.017	0.424	0.702	0.420	0.353
CISA (ours)	**0.737**	**0.533**	**0.022**	**0.601**	**0.863**	**0.569**	**0.490**	**0.714**	**0.426**	**0.014**	**0.477**	0.736	**0.442**	**0.392**

referring expression grounding, reasoning over sound semantics and commonsense knowledge, no-target rejection, or explanation generation, which are central to R2-AVCOS.

As shown in Tables 12 and 13, our model achieves competitive performance on both image-level and video-level COS benchmarks. These results indicate that the proposed model retains strong generic camouflaged object segmentation ability under conventional COS settings. More importantly, the standard COS results provide complementary evidence that our model's improvements on R2-AVCOS are not obtained at the cost of basic mask prediction ability, but are built upon a robust camouflaged segmentation foundation.

E Auxiliary Training Data Formulation (Additional)

Following Ref-AVS Bench [7], which studies referring audio-visual segmentation in natural scenes, we adopt an instruction-following formulation for audio-visual segmentation training. Ref-AVS provides a useful training paradigm for associating textual expressions and audio cues with pixel-level target masks. However, our R2-AVCOS task further extends this setting to camouflaged audio-visual scenes, where the model is required to understand sound semantics, visual camouflage cues, implicit reasoning expressions, no-target cases, and explanation-oriented outputs.

As summarized in Table 14, our auxiliary training data are drawn from the datasets used in our training setting. Specifically, we use semantic segmentation datasets, referring and reasoning image segmentation datasets, referring video segmentation datasets, audio-visual segmentation datasets, and audio-text alignment data. These auxiliary sources are used only for training and are complementary

Table 14: Auxiliary training data used in our method. "Mask", "Text", and "QA" denote pixel-level mask supervision, textual expressions/instructions, and question–answer supervision, respectively.

Dataset	Video	Audio	Mask	Text	QA	Role in Training
R2-AVCOSBench (Ours)	✓	✓	✓	✓	✓	In-domain audio-visual camouflaged segmentation and reasoning
Image-level segmentation data						
ADE20K [31]	✗	✗	✓	✗	✗	Generic semantic mask supervision
COCO-Stuff [32]	✗	✗	✓	✗	✗	Generic semantic mask supervision
PASCAL-Part [33]	✗	✗	✓	✗	✗	Part-level mask supervision
PACO-LVIS [34]	✗	✗	✓	✗	✗	Fine-grained mask supervision
Referring and reasoning image segmentation data						
RefCOCO [35]	✗	✗	✓	✓	✗	Image-level language grounding
RefCOCO+ [35]	✗	✗	✓	✓	✗	Image-level language grounding
RefCOCOg [36]	✗	✗	✓	✓	✗	Image-level language grounding
ReasonSeg [17]	✗	✗	✓	✓	✓	Reasoning-based image segmentation
Referring video segmentation data						
Refer-YouTube-VOS [37]	✓	✗	✓	✓	✗	Temporal referring segmentation
Refer-DAVIS-17 [38]	✓	✗	✓	✓	✗	Temporal referring segmentation
MeViS [39]	✓	✗	✓	✓	✗	Motion-aware video grounding
ReVOS [20]	✓	✗	✓	✓	✗	Reasoning-based video segmentation
Audio-related training data						
Ref-AVS Bench [7]	✓	✓	✓	✓	✗	Audio-visual target association
ASR datasets [43]	✗	✓	✗	✓	✓	Speech-to-text audio-language alignment
Audio caption datasets [44]	✗	✓	✗	✓	✓	Audio captioning and sound-event understanding

to the in-domain supervision from the training split of R2-AVCOSBench. No test annotations from R2-AVCOSBench or any evaluation benchmark are used during training.

For image-level segmentation data, we follow the data formulation of LISA [17], where different segmentation sources are reformulated into a unified instruction-following format with binary-mask supervision. For video and audio-visual segmentation data, we follow the Ref-AVS-style formulation and extend it to frame-level mask prediction under textual and audio-visual conditions. For audio-text data, we use automatic speech recognition and audio captioning supervision to align audio representations with the LLM-compatible semantic space.

Semantic segmentation data. Semantic segmentation datasets provide images with pixel-level annotations. Following LISA [17], we randomly sample one or several semantic categories or annotated regions from each image and convert their annotations into binary target masks. Each sample is reformulated into a QA-style segmentation instruction:

> USER: ⟨IMAGE⟩ Can you segment the {class name} in this image?
> ASSISTANT: Sure, it is ⟨SEG⟩.

where ⟨IMAGE⟩ denotes the image-token placeholder and ⟨SEG⟩ denotes the segmentation token whose hidden representation is decoded into a binary mask. The adopted datasets include ADE20K [31], COCO-Stuff [32], PASCAL-Part [33], and PACO-LVIS [34]. These data strengthen generic pixel-level mask prediction and help stabilize boundary localization.

Referring and reasoning image segmentation data. Referring segmentation datasets provide an image and a textual expression that describes the target object. Reasoning segmentation data further require understanding implicit descriptions and commonsense cues. Following the instruction-style formulation, each sample is converted into:

> USER: ⟨IMAGE⟩ Please segment the object described by this expression: {expression}.
> ASSISTANT: Sure, it is ⟨SEG⟩.

where {expression} is the provided referring or reasoning expression. The adopted datasets include RefCOCO, RefCOCO+ [35], RefCOCOg [36], and ReasonSeg [17]. These data improve language-guided visual grounding and enhance the model's ability to associate textual expressions with pixel-level regions.

Referring video segmentation data. Referring video segmentation datasets provide videos, textual expressions, and frame-wise target masks. We formulate these data as video-level instruction-following segmentation samples:

> USER: ⟨VIDEO⟩ Please segment the object described by this expression throughout the video: {expression}.
> ASSISTANT: Sure, it is ⟨SEG⟩.

where ⟨VIDEO⟩ denotes the visual token sequence extracted from sampled video frames, and ⟨SEG⟩ is decoded into frame-level binary masks. The adopted datasets include Refer-YouTube-VOS [37], Refer-DAVIS-17 [38], MeViS [39], and ReVOS [20]. These data improve temporal grounding and target consistency across frames.

Audio-visual segmentation data. Audio-visual segmentation datasets provide synchronized videos, audio signals, expressions, and pixel-level masks of sounding or expression-relevant objects. We reformulate them into audio-visual instruction-following samples:

> USER: ⟨VIDEO⟩ ⟨AUDIO⟩ Please segment the object described by this expression: {expression}.
> ASSISTANT: Sure, it is ⟨SEG⟩.

where ⟨AUDIO⟩ denotes the audio-token sequence. We adopt Ref-AVS Bench [7] as the audio-visual segmentation source. These data help the model associate sound semantics with visual targets and provide useful supervision for audio-conditioned mask prediction.

Audio-text alignment data. Since InternVL2 does not natively support audio input, we first conduct an audio-text alignment stage. Specifically, we use Automatic Speech Recognition datasets [43] and audio caption datasets [44] to align audio features with the LLM feature space. For ASR data, the audio input is paired with transcription-oriented prompts, such as:

> USER: ⟨AUDIO⟩ Please convert this speech into text.
> USER: ⟨AUDIO⟩ What is said in this speech clip?
> USER: ⟨AUDIO⟩ Please transcribe this audio.

The model is trained to generate the corresponding transcript. For audio caption data, the audio input is paired with caption-oriented prompts, such as:

> USER: ⟨AUDIO⟩ Please describe what you hear in this audio.
> USER: ⟨AUDIO⟩ What is happening in this audio clip?
> USER: ⟨AUDIO⟩ Describe the audio.

The model is trained to generate the corresponding audio caption. These audio-text data help align audio representations with the LLM-compatible semantic space and strengthen audio-content understanding.

In-domain R2-AVCOS instruction tuning. After audio-text alignment, we perform audio-visual instructed segmentation tuning using the training split of R2-AVCOSBench. Each sample contains a video, its corresponding audio, a textual expression, and a pixel-level mask annotation. We use a unified prompt template:

> USER: ⟨VIDEO⟩ ⟨AUDIO⟩ Please segment the object described by this expression: {expression}.
> ASSISTANT: Sure, it is ⟨SEG⟩.

For no-target expressions, where no object in the scene satisfies the given expression, the response is formulated as:

> USER: ⟨VIDEO⟩ ⟨AUDIO⟩ Please segment the object described by this expression: {expression}.
> ASSISTANT: No target matches this expression.

For reasoning expressions with explanation annotations, the response is augmented with an explanation after the segmentation prediction:

Table 15: Stepwise analysis of auxiliary training data types on the **All Expressions** split of R2-AVCOSBench. We progressively add audio-text data, semantic/part-level segmentation data, referring image segmentation data, referring video segmentation data, and Ref-AVS Bench to the in-domain R2-AVCOS training data. METEOR is used to evaluate generated explanations.

ID	Setting	R2-AVCOS Data	Audio-Text Data	Semantic/Part Segmentation	Referring Image Segmentation	Referring Video Segmentation	Ref-AVS Bench	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	METEOR
1	R2-AVCOS only	✓	✗	✗	✗	✗	✗	39.0	42.4	40.7	26.8
2	+ Audio-Text Data	✓	✓	✗	✗	✗	✗	40.2	43.7	42.0	28.1
3	+ Semantic/Part	✓	✓	✓	✗	✗	✗	41.5	45.0	43.3	28.9
4	+ Referring Image	✓	✓	✓	✓	✗	✗	42.4	45.8	44.1	30.1
5	+ Referring Video	✓	✓	✓	✓	✓	✗	43.2	46.5	44.9	31.0
6	Full Setting	✓	✓	✓	✓	✓	✓	**44.0**	**47.5**	**45.8**	**32.4**

USER: ⟨VIDEO⟩ ⟨AUDIO⟩ Please segment the object described by this expression and explain your decision: {expression}.
ASSISTANT: Sure, it is ⟨SEG⟩. {explanation}

This formulation enables the model to jointly learn audio-visual grounding, intention-aware camouflaged target segmentation, no-target rejection, and explanation generation.

Contribution of Training Data Types. Table 15 reports the stepwise analysis of auxiliary training data types on the All Expressions split of R2-AVCOSBench. Starting from R2-AVCOS-only training, progressively adding each data source brings consistent improvements, showing that these auxiliary data provide complementary supervision for audio-visual camouflaged object segmentation. Introducing audio-text data first improves the overall performance, indicating that Automatic Speech Recognition and audio caption data help align audio representations with the LLM-compatible semantic space and strengthen audio-content understanding. Adding semantic segmentation data further improves mask quality, suggesting that dense pixel-level supervision helps stabilize boundary prediction and improve local completeness in camouflaged scenes. Referring image segmentation data provide additional gains by enhancing language-guided visual grounding, while referring video segmentation data further improve temporal target consistency across frames. Finally, incorporating Ref-AVS Bench brings the best performance, demonstrating the importance of audio-visual segmentation supervision for associating sound semantics, textual expressions, and pixel-level target masks. Overall, the full training configuration achieves the best results, validating that audio-text alignment, generic segmentation supervision, language grounding, temporal grounding, and audio-visual association are jointly beneficial for R2-AVCOS.

F Annotation Reliability and Quality Control

To ensure the reliability of R2-AVCOSBench, we adopt a multi-stage annotation and verification process, including video filtering, expression annotation, mask annotation, no-target validation, and explanation checking. We first reviewed 10,871 candidate videos and selected 2,654 videos with clear audio-visual correspondence, identifiable camouflaged targets, and sufficient scene complexity for referring or reasoning expressions. The final benchmark contains 21,232 annotated frames and 32,329 expressions, including 17,543 referring expressions and 14,786 reasoning expressions, covering 45 target categories. To increase target ambiguity, 78% of videos contain two or more camouflaged objects, and 32% contain three or more objects.

For expression annotation, each video is annotated with both referring and reasoning expressions. Referring expressions identify targets using explicit audio, visual, spatial, or temporal cues, while reasoning expressions infer targets from sound semantics, object behavior, scene context, camouflage cues, or commonsense knowledge. Each expression is checked to ensure that it is grounded in the video and corresponds to a clear target, unless it is designed as a no-target expression. Ambiguous, conflicting, or non-unique expressions are revised or removed. For reasoning expressions, we further provide explanations describing the key evidence supporting the target selection. For mask annotation, annotators segment the camouflaged objects indicated by each expression. Using the first-appearance bounding box as a reference, they preview the video, confirm the target, and annotate pixel-level masks across sampled frames. For multi-target expressions, all matched objects are annotated. For no-target expressions, no mask is assigned. All masks are manually checked and refined, especially in low-contrast regions where camouflaged targets blend into the background.

After initial annotation, we conduct manual quality control from four aspects: expression–target consistency, reasoning validity, no-target correctness, and mask boundary quality. Reviewers check whether the target matches the expression, whether the reasoning is supported by audio-visual evidence or commonsense, whether no-target cases are valid, and whether masks accurately cover the target across frames. Samples with inconsistent target assignment, unsupported reasoning, ambiguous no-target status, or inaccurate masks are revised through manual rechecking. This process ensures that R2-AVCOSBench is constructed not only at scale, but also with explicit manual verification, supporting evaluation under referring expressions, reasoning expressions, multi-object ambiguity, no-target cases, and explanation-oriented reasoning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction accurately summarize the paper's main contributions, including the R2-AVCOS task, the R2-AVCOSBench benchmark, and the proposed CISA model. The claims are supported by the main experiments, ablation studies, auxiliary COS evaluations, and dataset analyses in the main paper and Appendix B–F.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper explicitly discusses limitations in Section 6, *Conclusion and Discussion*. It notes that CISA still faces challenges in complex scenes with multiple sound sources and highly concealed objects, where disentangling audio events and associating them with correct camouflaged targets remains difficult, and suggests future directions such as stronger audio-visual fusion, reasoning mechanisms, larger-scale pre-training, and interactive settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: The paper does not present new theoretical results, formal theorems, or proofs. The main contributions are a new benchmark, a task formulation, a model design, and empirical evaluations.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper discloses the task formulation, dataset construction pipeline, annotation protocol, training data composition, prompt templates, evaluation metrics, and implementation details needed to reproduce the main experiments. The CISA code and a demo dataset are released at submission, and the full R2-AVCOSBench dataset will be released after acceptance due to its large scale.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The paper provides detailed implementation settings, training configurations, evaluation protocols, and dataset construction procedures to support reproducibility. The code and dataset will be released together with the finalized public version of this project.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The paper specifies the training and evaluation settings, including benchmark splits, auxiliary training datasets, prompt formulations, evaluation metrics, and the role of each training source. Additional details are provided in the experimental section and Appendix D–E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No].

Justification: The paper does not report error bars or statistical significance tests for all experiments because the evaluation involves large-scale video segmentation, multiple baselines, and high computational cost. Instead, the paper provides comprehensive benchmark comparisons, ablations, auxiliary evaluations, and consistent metric-based analysis.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The paper reports the main computational resources and implementation environment used for training and evaluation, including the GPU setting and training configuration. These details are provided in the implementation details section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The research follows the NeurIPS Code of Ethics. The benchmark is constructed with manual verification, no-target validation, mask quality control, and explanation checking to reduce annotation ambiguity and unsupported reasoning, as described in Appendix F.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper discusses the positive impact of improving audio-visual reasoning and camouflaged target understanding, as well as potential risks related to misuse in surveillance-like scenarios. The discussion emphasizes responsible research use and controlled release of the benchmark.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: This work does not release high-risk models such as pre-trained language models or image generators. The current submission also does not publicly release the complete benchmark dataset or model weights. Therefore, safeguards for the public release of high-risk data or models are not applicable to this submission.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper properly cites the existing datasets, models, and benchmarks used for training and evaluation, including semantic segmentation, referring segmentation, video segmentation, audio-visual segmentation, ASR, and audio caption datasets. The auxiliary training sources are summarized in Appendix E, and the use of existing assets follows their licenses and terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper introduces R2-AVCOSBench as a new benchmark and provides detailed documentation of its construction process, data scale, expression types, mask annotations, no-target cases, explanation annotations, evaluation protocols, and quality-control process. The complete benchmark and accompanying documentation will be organized and released with the finalized public version of the project.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Human involvement is limited to expert annotation and quality verification for dataset construction, and the annotation and verification procedures are described in Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not conduct behavioral experiments or collect private information from human subjects. The human involvement is limited to dataset annotation and verification, and annotators are not studied as research subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes].

Justification: The paper uses a pretrained multimodal large language model as part of the core model architecture and describes how audio representations are aligned with the LLM-compatible semantic space. The audio-text alignment stage, instruction-following formulation, and segmentation response format are described in the method section and Appendix E.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.