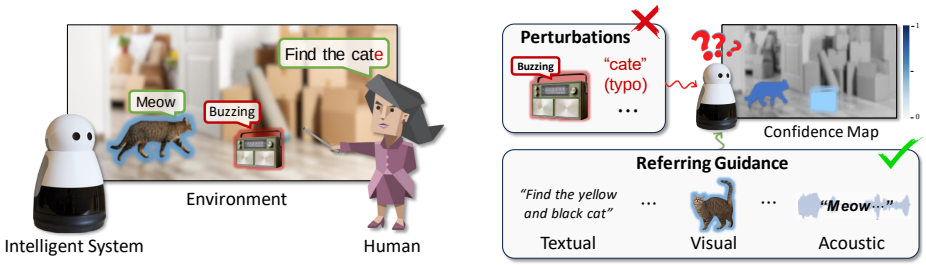
R^2-Bench: Benchmarking the Robustness of Referring Perception Models under Perturbations

Xiang Li[1], Kai Qiu[1], Jinglu Wang[2], Xiaohao Xu[3], Rita Singh[1], Kashu Yamazaki[1], Hao Chen[1], Xiaonan Huang[3], Bhiksha Raj[1,4]

[1] CMU, [2] Microsoft Research Asia, [3] University of Michigan, [4] MBZUAI



(a) Referring Perception in Practice

(b) Perturbation in Referring Guidance

Fig. 1: Motivation illustration. Referring perception models (RPMs) empower intelligent systems with their ability to perform object grounding within the environment based on referring guidance, such as textual descriptions, imagery exemplars, or auditory signals associated with the target object. However, RPMs' performance can be compromised by disturbances in real-world scenarios, such as environmental noise (*e.g.*, extraneous sounds from a nearby radio), human-induced errors (*e.g.*, typographical errors in textual input), and limitations in the sensor (*e.g.*, motion blur in images). Conducting a rigorous analysis of RPMs' robustness to a wide array of perturbations is necessary for building reliable real-world applications.

Abstract. Referring perception, which aims at grounding visual objects with multimodal referring guidance, is essential for bridging the gap between humans, who provide instructions, and the environment where intelligent systems perceive. Despite progress in this field, the robustness of referring perception models (RPMs) against disruptive perturbations is not well explored. This work thoroughly assesses the resilience of RPMs against various perturbations in both general and specific contexts. Recognizing the complex nature of referring perception tasks, we present a comprehensive taxonomy of perturbations, and then develop a versatile toolbox for synthesizing and evaluating the effects of composite disturbances. Employing this toolbox, we construct **R^2-Bench**, a benchmark for assessing the **R**obustness of **R**eferring perception models under noisy conditions across five key tasks. Moreover, we propose the **R^2-Agent**, an LLM-based agent that simplifies and automates model evaluation via natural language instructions. Our investigation uncovers the vulnerabilities of current RPMs to various perturbations and provides tools for assessing model robustness, potentially promoting the safe and resilient integration of intelligent systems into complex real-world scenarios.

Keywords: Referring perception · Robustness & Perturbation · Benchmark

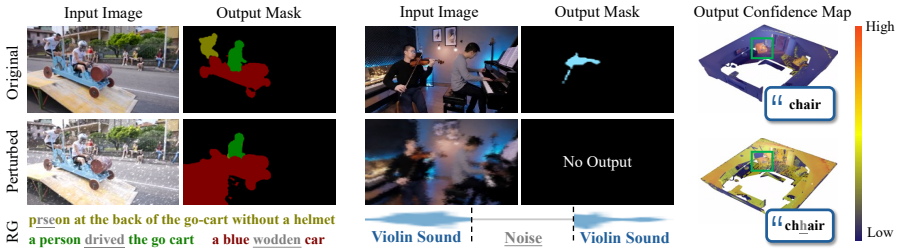


Fig. 2: Examples of R^2-Bench. The "Original" row displays the original inputs alongside the outcomes from models of R-VOS [54], AVS [34] and Q3M [23], while the "Perturbed" row presents the inputs as synthesized by R^2-Bench and the respective outcomes with the same models. RG: short for Referring Guidance.

1 Introduction

Perception systems [6–8,29–32,80,81] function as input channels for intelligent systems, analogous to human eyes. *Referring perception*, focuses on identification and contextualization of visual entities by referring multimodal guidance, illustrated in Fig. 1 (a). It effectively creates a communicative bridge between humans, who issue instructions, and the environment that is subject to perception. Multimodal referring cues used for such identification include textual descriptions, example imagery, or auditory signals corresponding to the target object (Fig. 1 (b)). Within the domain of referring perception, specific tasks such as referring expression segmentation [37,68], audiovisual source localization [17,34,79], and queryable 3D mapping [23], are essential modules of robotic control [2,22], autonomous navigation [23,40] and planning [21,48], and human-computer interaction [51].

Recent advancements in referring perception models [14,33,68] have been witnessed and deployed on resource-constrained platforms [54,58]. However, existing datasets for model training often comprise delicate images with accurate annotations, a condition rarely met in real-world scenarios. In realistic settings, as illustrated in Fig. 1, perception systems may face various disturbances, such as environmental noise (buzzing sound from the radio), imprecise human instructions (misspelling), and sensor imperfections (image blur), which can significantly challenge the robustness of these models. Despite their critical importance for safer real-world applications, the impact of such perturbations on model performance is not thoroughly investigated in the literature. Evaluating model robustness is further complicated by the diversity of operational environments; the assessment process is typically labor-intensive and relies heavily on expert judgment. Addressing this gap, this paper introduces the first benchmark in the field to systematically assess the resilience of referring perception models, *i.e.*, trying to obtain a reliable and comprehensive answer to the critical question: "*How robust are the current referring perception models under realistic perturbations?*"

Specifically, we introduce the **R**obust **R**eferring Benchmark (**R^2-Bench**), featuring: (1) a human-friendly taxonomy of perturbations for referring perception,

categorizing disruptions into source, environment, sensor, and transmission noise, and detailing perturbation types for each modality; (2) a customizable perturbation synthesis toolbox for creating reasonable noise-augmented datasets, enabling robustness evaluation of perception models; and (3) robustness evaluations across five tasks—referring image segmentation (RIS), video object segmentation (VOS), referring video object segmentation (R-VOS), audiovisual segmentation (AVS), and queryable 3D mapping (Q3M)—considering textual, visual, and acoustic reference modalities. We assess over twenty prevalent models in noisy conditions to benchmark their performance and analyze their vulnerabilities. (Examples are illustrated in Fig. 2.) We hope our benchmarking and analysis can benefit the community by shedding light on the vulnerability of existing models in the face of various perturbations.

Notably, our work advances the evaluation of model robustness by incorporating tests against composite noise types, more closely mirroring real-world scenarios where different noises often coexist. The construction of complex noise for realistic robustness assessment is not trivial, as it must avoid non-contextual perturbations, such as the presence of snow in indoor imagery or excessive air absorption effects in indoor audio recordings, which are incongruent with the depicted environment. To address this complexity, we introduce the **R^2-Agent**, a novel large language model-based (LLM-based) system that can comprehend natural language instructions provided by humans and autonomously generates test samples with contextual relevant perturbations that align with the deployment environment. For instance, given the human-provided instruction "outdoor night scene", the R^2-Agent could probably generate perturbations of "motion blur" instead of "glass blur" and "low background noise" instead of "room reverberation". To improve the agent's abilities of instruction-following and commonsense-reasoning, we employ the multi-agent debating technique [16,36,77], wherein one agent proposes potential solutions while another verifies their validity. Consequently, the R^2-Agent efficiently automates the identification and integration of mixed perturbations, thus optimizing the process of robustness evaluations tailored to specific domain contexts.

In summary, our major contribution is three-fold:

- We introduce the R^2-Bench, a benchmarking framework that includes diverse perturbed data for five commonly encountered referring perception tasks. A comprehensive taxonomy and a corresponding customizable synthesis toolbox are developed to enable robustness assessment in noisy real-world settings.
- We propose the R^2-Agent to streamline model evaluations tailored to particular use scenarios based on LLMs. This automated program executes the perturbation composition process in accordance with human-provided instructions, thereby allowing for context-specific robustness evaluations.
- Our systematic experimental investigations delve into the intrinsic characteristics of perturbations, such as their types, severity, dynamics, and correlations. These explorations yield valuable insights into the vulnerability of existing models to disturbances and elucidate the nature of these perturbations.

2 Related Works

Textual referring perception. Referring image segmentation (RIS) and referring video object segmentation (R-VOS) aim to segment objects in images and video sequences, respectively, based on a linguistic description. Recent RIS methods [15, 37, 63, 68, 82–84] have achieved promising results using multimodal transformers. R-VOS [14, 18, 42, 47, 49, 54] is more challenging as it requires leveraging both intra-frame and temporal cues. URVOS [47] is the first unified R-VOS framework with cross-modal attention and a memory attention module, significantly improving R-VOS performance. ReferFormer [55] employs a linguistic prior in the transformer decoder to focus on the referred object, while MTTR [5] uses a multimodal transformer encoder to fuse linguistic and visual features. R^2-VOS introduces relational cyclic consistency to enhance the robustness of the R-VOS model. Unlike other vision-language tasks [65, 66], R-VOS needs to construct object-level multimodal semantic consensus in dense visual representations. Relying on 2D models, 3D referring perception is also achievable [23, 64].

Acoustic referring perception. Audiovisual segmentation (AVS) [35, 35, 38, 78] focuses on segmenting objects that produce sound in a given image frame. This pioneering work by Zhou et al. [78] introduced a method that uses cross-modal attention to identify the sound source. Building on this, Zhou et al. [78] proposed an extended task called audiovisual semantic segmentation (AVSS), which not only segments the sound-producing objects but also classifies them. AVSS is more challenging than AVS due to the complexity of audio semantics. To address this, Zhou et al. [78] utilized the TPAVI module from [79] for audiovisual interaction. Additionally, a recent development by Li et al. [27] introduced the CATR framework, which features a novel spatial-temporal audio-video fusion block for efficient audio-visual integration. Li et al. [34] propose a quantization-based semantic decomposition module to handle the complex acoustic environment and enhance the robustness. Sound source localization (SSL) is a related field that aims to identify the visual regions corresponding to sounds. Several SSL methods [3, 4, 12, 46] utilize the correspondence between audio and visual features to locate sounds, often represented as heatmaps. For example, Mo et al. [43] employed multi-level audiovisual contrastive learning to effectively pinpoint sound-producing objects. In addition, using speech as referring guidance is also explored in [35, 44]

Visual referring perception. Visual referring perception typically relies on visual prompts such as example images, points, bounding boxes, or scribbles. Segment anything model (SAM) [26] is a powerful foundation model that supports multiple visual prompts. Several follow-up works [24, 39, 58] improve SAM in terms of inference cost, segmentation quality, and referring prompts. Beyond that, several works [67, 73, 75] explore the usage of SAM in downstream tasks. Li et al. [30] built a SAM-like model with stable diffusion. For video-level tasks, semi-video object segmentation (VOS) [9, 28, 60, 70, 71] aims to segment visual objects across frames given the first frame annotation. Recent offline models [9, 11] focus more

on designing long-term information propagation modules to transfer previous image features and corresponding masks to the target frame to predict masks. This helps more precisely identify and track the movement trajectory of objects throughout the entire video sequence (more than 200 frames) but also makes the network architecture heavy [72]. Due to the slow speed of offline models, online models [56, 76] come back into focus, aiming to keep the right balance between speed and performance.

3 R^2-Bench: Customizable Perturbation Benchmark

In this section, we present the R^2-Bench for evaluating robustness of referring perception tasks, including a comprehensive taxonomy that categorizes various perturbations and a customizable perturbation synthesis toolbox that can be tailored to generate specific disturbances.

3.1 Taxonomic Perturbations in Referring Perception

Referring perception tasks typically employ multimodal guidance, spanning the visual, auditory, and textural domains, to ground objects in the visual contexts. Nonetheless, in real-world scenarios, each modality may undergo heterogeneous noises before input into the model. As depicted in Fig. 3, noises can be categorized into four classes based on their origins: source noises, environment noises, sensor noises, and transmission noises.

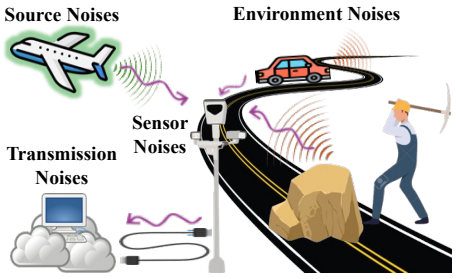


Fig. 3: Noise categories based on their origins. Assuming airplane as the source of referring guidance, noise from it is categorized as source noise.

Source noises. We define source noise as the inherent disturbances introduced spontaneously by the source of the referring guidance. For example, for text-referring tasks, the source noises can manifest as misspellings, grammar errors, or punctuation errors in the textual expression and instruction (human is usually the source of text reference).

Environment noises. We define environment noise as the unintended disturbance introduced by the surrounding environment during the signal capture process. Background sound, room reverberation and air absorption for acoustic waveforms are common examples of this type.

Sensor noises. We define sensor noise as the unwanted disturbances in the output of sensors that arise from inherent limitations or imperfections. For example, defocus blur, motion blur and impulse noise for visual frames.

Table 1: Noise types considered in R^2-Bench. Details are available in the Appendix.

Type	Source	Environment	Sensor	Transmission
Visual	-	snow (SN), fog (FG), frost (FT), spatter (SP), brightness (BR)	defocus blur (DB), gaussian blur (GB), motion blur (MB), glass blur (GS), impulse noise (IN), shot noise (ST), speckle noise (SPN), contrast (CT), saturate (SA)	JPEG compression (JPG), pixelated (PIX)
Acoustic	amplitude gain (GA)	background noise (BN), air absorption (AA), room reverberation (RS)	gaussian noise (GN), impulse noise (IN), peak filter (PF), time mask (TM), tanh distortion (TD)	MP3 compression (MP3), lowpass filter (LP), highpass filter (HP)
Textual	misspelling (MS), mispunctuation (MP), grammar error (GE)	-	character missing (CM)	-

Transmission noises. We define transmission noise as the signal distortion or loss during the transmission resulting from data compression or package loss. For example, JPEG and MP4 compression for visual and acoustic signals.

We summarize the supported noises in visual, acoustic and textual modalities in Tab. 1. A total of 32 types of noises are considered in this paper. The implementation details of noise functions are available in the Appendix.

3.2 Customizable Perturbation Synthesis

We approach the robustness evaluation through two conventional paradigms: (1) general evaluation under universal scenarios and (2) specific evaluation under designated instructions. To facilitate these evaluations, it is essential to generate perturbed datasets from the original clean data samples, denoted as $\mathcal{X} = \{x\}$, by introducing a spectrum of heterogeneous perturbations $\Delta = \{\delta_k\}$.

Perturbation order. As discussed in Sec. 3.1, perturbations arise sequentially according to their origins in real-world contexts. To emulate this, perturbations are applied in a sequence that mirrors the natural order of disruption: source→environment→sensor→transmission. Within a single category of origin, the order is randomly determined to account for the non-deterministic nature of real-world noise. To formalize the perturbation process, consider a clean data sample denoted by x, the perturbed sample x' undergoes a series of transformations that can be mathematically described as follows:

$$x' = \delta_t \circ \delta_{se} \circ \delta_e \circ \delta_{so}(x), \tag{1}$$

where $\circ$ denotes composition operation. $\delta_{so}, \delta_e, \delta_{se}, \delta_t$ represent generation functions for source, environment, sensor, and transmission noises, respectively. Given that the introduction of perturbations follows a sequential order, the operations that compose these noise functions are inherently *order-variant*.

Perturbation severity and mode. In alignment with established methodologies [20, 61, 62], we quantitatively define perturbation severity across discrete levels that align with human perceptual categories, namely, {low, medium, high}. In addition, we investigate perturbation across two main modes based on dynamics:

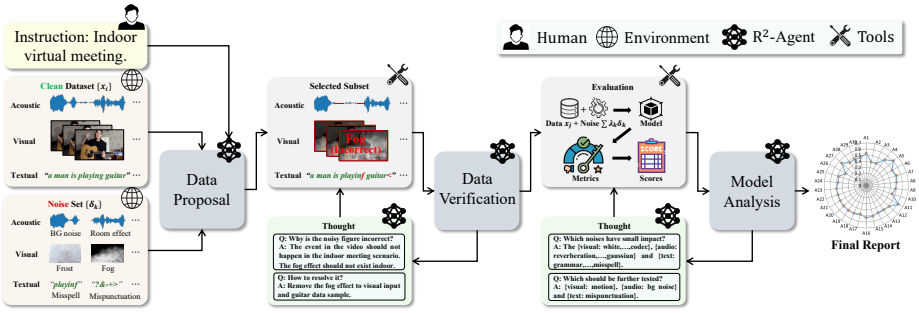


Fig. 4: Overview of R^2-Agent, the automatic evaluation assistant. Given a human instruction, clean datasets, perturbation functions, and evaluation functions, R^2-Agent first proposes and verifies perturbed test samples that match the given instruction. After that, R^2-Agent evaluates the model using the verified samples and provides a report that articulates the model's vulnerabilities and overall resilience.

{static, dynamic}. Static perturbations remain a consistent level of severity throughout all sensor frames within a sequence, while dynamic perturbations exhibit varying severity and types from frame to frame, closely simulating the fluctuating nature of disturbances in real-world environments, taking the form:

$$x' = \circ_{k=1}^{K} \delta_k(x, \lambda_k(t)), \tag{2}$$

where $\lambda_k(t)$ denotes the severity of the k-th perturbation at time step t.

For the general evaluation under universal scenarios, we construct perturbed datasets for each task by leveraging the original clean benchmarks and all perturbations. Specifically, noisy datasets with low, medium, high and dynamic severity are considered. In each setting, a maximum of two perturbations are considered simultaneously. The detailed data-creating procedure is available in the Appendix. Comprehensive experiments and analyses discussing the impact of perturbations are provided to facilitate future research in Sec. 5.3.

Furthermore, we also consider specific evaluations under designated instructions in practice, *e.g.*, virtual phone meetings without background noise. We detail R^2-Agent in Sec. 4.

4 R^2-Agent: Automatic Evaluation Assistant

With the versatile capabilities of LLMs [1, 50], LLM-based agents [52, 57] are employed to automatically perform various tasks in a human-like manner. In this section, we introduce R^2-Agent, an LLM-based agent that facilitates automatic perturbation synthesis and robustness evaluation guided by human instructions. As depicted in Fig. 4, the R^2-Agent consists of three primary steps: data proposal, data verification, and model analysis. We utilize Gemini-Vision-Pro [50] as the LLM component within the R^2-Agent.

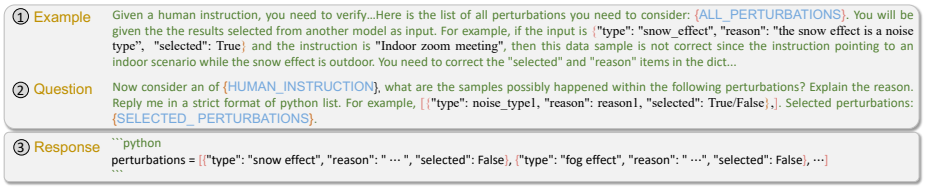


Fig. 5: Chain-of-thought prompting template for data verification. ①&②: Following the chain-of-thought spirit, we first give examples to LLM to boost the in-context-learning capability. After that, we ask the LLM to answer a question that is similar to the given example. Specifically, for the data verification task, we ask the LLM to verify the selected samples from the previous iteration, update the results, and explain the reason. The LLM response is instructed to be a Python-format list of dictionaries. ③: The response from LLM (which follows the desired Python-format list).

4.1 Evaluation Pipeline

Given a set of clean data samples $\mathcal{X} = \{x_i\}$, perturbation functions $\Delta = \{\delta_k\}$, evaluation metric functions $\mathcal{E} = \{e_j\}$, and a human instruction u, the objective of R^2-Agent is to select an appropriate subset of clean data samples $\mathcal{X}_u$, potential perturbations Δ_u, and suitable evaluation functions $\mathcal{E}_u$ to assess the model based on the given instruction u. A desired human instruction for the R^2-Agent should explicitly specify the scenario (e.g., indoor), the evaluation focus (e.g., segmentation quality), and, if applicable, any specific requirements (e.g., testing for motion blur) during the evaluation process. To enhance the robustness of the R^2-Agent, we introduce a multi-agent debating strategy utilizing an *operator* agent Φ and a *guardian* agent Ψ, each with their own independent memories.

Data proposal. To propose data samples that align with the criteria delineated in the human instruction, we employ the *operator* agent Φ with the text prompts `<SEL>`, Φ enables the traversal through the combined space of clean samples, pre-defined noise generation functions, and evaluation functions, collectively represented by the tuple $(\mathcal{X}, \Delta, \mathcal{E})$, to select suitable samples. Additionally, we employ an LLM to generate captions $\mathcal{C}$ for the clean samples $\mathcal{X}$. These captions are archived within an accessible memory to enhance the computational efficiency for subsequent stages. The data proposal stage can be formulated as:

$$\mathcal{C}, \mathcal{X}_u, \mathcal{E}_u, \Delta_u = \Phi(u, \mathcal{X}, \mathcal{E}, \Delta, \texttt{<SEL>}). \tag{3}$$

Data verification. Although LLMs exhibit notable proficiency in reasoning tasks, their susceptibility to generating "hallucinated" content poses a significant challenge to the delivery of accurate responses. To mitigate the hallucination effects, we introduce a verification mechanism, designated as the safe guardian Ψ, which monitors the data proposal made by the operator Φ through the utilization of chain-of-thought prompting [53]. We first create the proposed dataset $\mathcal{X}_u^t$ and perturbations Δ_u^t. t denotes the iteration number. Then, we construct verification prompts `<VER>` and employ Ψ to validate the integrity of the data instances. Feedback from this verification is relayed to Φ, which then uses the information

Table 2: Summary of considered tasks, metrics, referring modality, and datasets.

Task	RIS	VOS	R-VOS	AVS	Q3M
Ref. Modality	Text	Image	Text	Audio	Text
Metrics	mIoU	$\mathcal{G}$, $\mathcal{J}$, $\mathcal{F}$ [25]	$\mathcal{J}$, $\mathcal{F}$ [25]	$\mathcal{J}$, $\mathcal{F}$ [25]	-
Dataset	RefCOCO [74]/+ [74]/g [41]	DAVIS [25], YTVOS [59]	Ref-DAVIS [25], Ref-YTVOS [47]	AVS-s4 [78], AVS-ms3 [78]	ScanNet [13], ICL [19]

to refine and enhance the data proposal. This iterative process is carried out until a state of convergence or until a maximum iteration number is reached, which is formulated as:

$$\begin{aligned} R^t &= \Psi(\mathcal{X}_u^t, \Delta_u^t, \texttt{<VER>}_\Psi) \\ \mathcal{X}_u^{t+1}, \Delta_u^{t+1} &= \Phi(R^t, \texttt{<VER>}_\Phi), \end{aligned} \tag{4}$$

where t is the iteration index, R^t is the response of Ψ at t iteration. Note that we consider Φ and Ψ have separate memories for previous inputs (similar to two separate ChatGPT sessions) and we omit the memorized inputs here for simplicity.

Specifically, adversarial prompts are leveraged in the data verification step. In $\texttt{<VER>}_\Phi$, we ask the LLM to select conservative choices with high confidence while, for $\texttt{<VER>}_\Psi$, radical choices are acceptable. The adversarial prompts can proactively encourage a debate between two agents thus correcting potential mistakes in the original data proposal. To help understand the data verification step, we demonstrate an example of the data verification process in Fig. 5.

Model analysis. With the selected samples and functions $\mathcal{X}_u, \mathcal{E}_u, \Delta_u$, we instantiate the noisy data $\mathcal{X}_u'$ by applying perturbation functions to the clean data, $\mathcal{X}_u' = \Delta_u(\mathcal{X}_u)$. We then calculate a set of metrics $\mathcal{M}_u = \{m_i\}$ that correspond to $\mathcal{X}_u'$ by applying evaluating functions $\mathcal{E}_u$ to noisy samples. Utilizing the chain-of-though prompt <ANA>, we engage an LLM Θ with empty memory to analyze the metrics and produce an output report O for the model performance as:

$$O = \Theta(\mathcal{M}_u, \mathcal{X}_u', \texttt{<ANA>}). \tag{5}$$

5 Experiment

5.1 Evaluation Setup and Metrics

In Tab. 2, we present a summary of the benchmark tasks, evaluated datasets, and corresponding metrics. Our benchmark encompasses five prevalent tasks, namely, RIS, VOS, R-VOS, AVS and Q3M.

Evaluation setup. We conduct systematic evaluations to analyze the impact of perturbations on state-of-the-art mode five tasks. Both perturbations in the referring guidance and visual frames are considered in the benchmarking. We create noisy datasets with low, medium and high noise levels for each task. For video-level tasks, we additionally consider the scenario that noises change over time. We first benchmark the models' general performance with all types of noises and then investigate the impact of noises individually. More information about the dataset creation is available in the Appendix.

Table 3: Performance of referring image segmentation methods under low, medium, and high perturbation levels. Average performance change (APC) is averaged among all perturbation levels. We report mIoU following convention.

Method	Venue	RefCOCO					RefCOCO+					RefCOCOg				
		Clean	Low	Med.	High	APC	Clean	Low	Med.	High	APC	Clean	Low	Med.	High	APC
LAVT [69]	CVPR 22	0.74	0.58	0.53	0.48	-0.21	0.66	0.58	0.51	0.43	-0.15	0.63	0.55	0.47	0.38	-0.16
PolyFormer [37]	CVPR 23	**0.76**	**0.70**	**0.64**	**0.57**	**-0.12**	**0.71**	**0.65**	**0.57**	**0.50**	-0.14	**0.69**	**0.64**	**0.57**	**0.50**	**-0.12**
X-Decoder [82]	CVPR 23	0.68	0.60	0.52	0.44	-0.16	0.69	0.60	0.51	0.44	-0.17	0.64	0.58	0.50	0.44	-0.14
ETRIS [63]	ICCV 23	0.71	0.64	0.57	0.49	-0.14	0.60	0.56	0.48	0.40	**-0.12**	0.60	0.53	0.45	0.38	-0.15
SEEM [84]	NeurIPS 23	0.65	0.59	0.54	0.47	**-0.12**	0.64	0.58	0.51	0.44	-0.13	0.62	0.56	0.49	0.43	**-0.12**

Table 4: Performance of video object segmentation methods under low, medium, and high perturbation levels. APC is averaged among all perturbation levels.

Method	Venue	DAVIS										YTVOS									
		Clean		Low		Medium		High		APC		Clean		Low		Medium		High		APC	
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}_s$	$\mathcal{G}_u$	$\mathcal{G}_s$	$\mathcal{G}_u$	$\mathcal{G}_s$	$\mathcal{G}_u$	$\mathcal{G}_s$	$\mathcal{G}_u$	$\mathcal{G}_s$	$\mathcal{G}_u$
AOT [70]	NeurIPS 21	76.0	83.0	71.7	79.2	69.7	76.9	65.3	71.1	-7.1	-7.3	86.4	84.2	84.9	80.9	82.0	78.1	76.1	70.4	-5.4	-7.7
DeAOT [71]	NeurIPS 22	76.9	84.5	73.2	81.0	72.4	80.4	68.2	75.5	-5.6	-6.4	86.7	**84.5**	86.6	**83.7**	84.5	**81.8**	**80.8**	**76.8**	**-2.7**	**-3.7**
XMem [11]	ECCV 22	77.4	84.5	72.2	80.0	71.8	79.7	68.4	74.7	-6.6	-6.4	86.5	**84.5**	84.0	81.2	81.7	79.0	78.6	74.9	-5.1	-6.1
DEVA [10]	ICCV 23	78.7	85.9	**76.4**	**84.9**	74.2	82.4	**69.9**	**77.5**	**-5.2**	**-4.3**	87.2	83.9	**86.7**	82.8	**84.7**	80.0	80.6	76.3	-3.2	-4.2
Cutie [9]	CVPR 24	**80.6**	**87.7**	75.8	83.5	**75.0**	**82.7**	69.3	75.4	-7.2	-7.2	**87.8**	**84.5**	86.5	82.8	84.2	80.5	80.3	75.1	-4.1	-5.0

Table 5: Performance of referring video object segmentation methods under low, medium, and high perturbation levels. APC is averaged among all perturbation levels.

Method	Venue	Ref-DAVIS										Ref-YTVOS									
		Clean		Low		Medium		High		APC		Clean		Low		Medium		High		APC	
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Video-level Backbone																					
MTTR [5]	CVPR 22	-	-	-	-	-	-	-	-	-	-	54.0	56.4	48.9	51.6	43.0	45.5	38.0	40.7	-10.7	-10.5
SgMg [42]	ICCV 23	**59.0**	**64.8**	**57.3**	**62.6**	**51.9**	**57.7**	**47.3**	**51.1**	**-6.8**	**-7.7**	**57.7**	**60.0**	**57.4**	**59.9**	**52.4**	**54.5**	**44.1**	**46.2**	**-6.4**	**-6.5**
Image-level Backbone																					
ReferFormer [55]	CVPR 22	55.8	61.3	46.5	50.1	43.1	48.0	38.5	41.9	-13.1	-14.6	54.8	56.5	41.1	42.0	37.4	37.8	33.7	34.2	-17.4	-18.5
OnlineRefer [54]	ICCV 23	55.7	**62.9**	**51.6**	**57.9**	**48.0**	**54.3**	**42.2**	**46.3**	**-8.4**	**-10.0**	55.6	**58.9**	**51.5**	**54.3**	**47.6**	**49.8**	**39.6**	**42.0**	**-9.4**	-10.2
R^2-VOS [33]	ICCV 23	**57.2**	62.4	50.2	57.0	46.5	53.8	39.2	44.6	-11.9	-10.6	**56.1**	58.4	48.4	51.2	45.3	47.4	38.5	40.7	-12.0	**-9.7**

Metrics. Following the convention [41, 74], we leverage mIoU to evaluate the RIS task. For video-level tasks, the convention is to compute region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ as defined in [45]. Specifically, $\mathcal{G} = \frac{\mathcal{J}+\mathcal{F}}{2}$ is also used. In addition, to better demonstrate the performance degradation against perturbations, we define average performance change (APC) as $\text{APC} = \sum_i \frac{m_i^n - m_i^c}{N}$ where m_i^n and m_i^c denotes the performance of the i-th sample of perturbed and clean data respectively. N is the sample number. Additional results, evaluated using various metrics and settings, can be found in the Appendix.

5.2 Performance Benchmarking

Referring image segmentation. As shown in Tab. 3, we evaluate the state-of-the-art RIS methods on perturbed RefCOCO/+/g [74] datasets. We notice that PolyFormer [37] achieves the best performance in terms of both performance and robustness. Different from other methods that generate pixel-level classification results to represent an object mask, PolyFormer utilizes multi-polygon vertices to represent the mask. We consider the robustness can result from the special sequential prediction and polygon representation which eases the reliance on the per-pixel representation. In addition, we notice that though SEEM [83] shows an inferior performance, its robustness to perturbations is promising. We consider that multi-task joint training can account for SEEM's robustness.

Table 6: Performance of audiovisual segmentation methods under low, medium, and high perturbation levels. APC is averaged among all perturbation levels.

Method	Venue	AVS-s4										AVS-ms3									
		Clean		Low		Medium		High		APC		Clean		Low		Medium		High		APC	
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
AVS [79]	ECCV 22	72.8	84.8	68.4	79.7	60.8	71.2	55.6	66.8	-11.2	-12.2	47.9	57.8	44.4	54.6	41.2	50.5	37.0	44.2	-7.0	**-8.0**
CATR [27]	MM 23	74.9	**87.1**	71.1	**84.2**	65.7	**79.0**	58.4	**72.4**	**-9.8**	-8.6	53.1	**65.6**	49.8	**61.4**	46.7	**58.4**	41.2	**52.9**	-7.2	**-8.0**
AVSegFormer [17]	AAAI 23	76.4	86.7	70.9	82.3	62.4	74.8	57.3	69.8	-12.9	-11.0	49.5	62.8	46.9	59.0	43.7	55.4	39.8	50.1	**-6.0**	**-8.0**
QSD [35]	CVPR 24	**77.6**	85.6	**74.8**	83.6	**68.3**	78.4	**59.4**	71.2	-10.1	**-7.9**	**61.8**	64.3	**56.6**	60.0	**53.1**	57.1	**47.6**	51.6	-9.4	**-8.0**

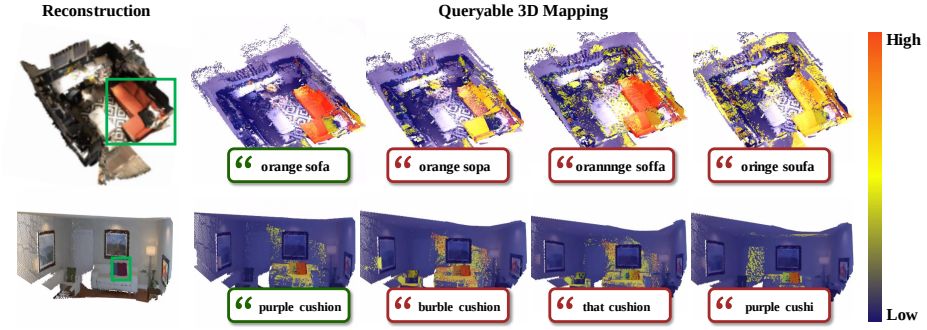


Fig. 6: Queryable 3D mapping of ConceptFusion [23] with inaccurate referring guidance on ScanNet [13] and ICL [19] datasets. Due to the absence of publicly available evaluation codes, we did not provide quantitative evaluation for queryable 3D mapping.

Video object segmentation. As depicted in Tab. 4, we demonstrate the performance of state-of-the-art methods on DAVIS [25] and YTVOS [59] datasets. We notice that, for simple scenarios in DAVIS datasets, Cutie [9] shows the best performance. While in the more complex YTVOS dataset, DeAOT [71] achieves the best. Cutie heavily relies on the pixel- and object-level correspondence across frames which can be disrupted by the perturbations. Differently, DeAOT leverages decoupled visual and ID embeddings which may be the reason for its robustness in noisy scenarios. In addition, we notice that the unseen categories generally have a larger performance drop compared to the seen categories which can impose more challenges in practical deployment in complex scenarios.

Referring video object segmentation. We benchmark the referring video object segmentation task in Tab. 5. SgMg [42] and OnlineFormer [54] achieve the best performance and robustness among methods equipped with video- and image-level backbones respectively. Unlike other methods, OnlineRefer processes the visual features in a frame-by-frame manner which makes it rely less on the pixel-level temporal correspondence that is easily been disrupted by visual perturbations.

Audio-visual segmentation. We show the performance of popular AVS methods in Tab. 6. We notice that QSD [34] and CATR [27] archive promising performance and robustness among all methods. CATR [27] and AVSegFormer [17] are similar methods that leverage acoustic query to query the visual frames with a transformer-based structure. QSD moves one step forward which additionally introduces a quantization operation and a local-global distillation to enhance the robustness of the acoustic representation to adapt to complex scenarios.

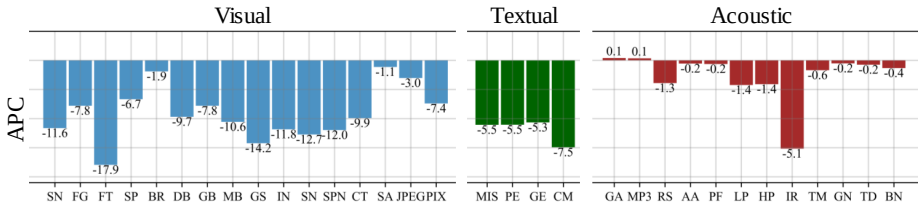


Fig. 7: Average performance changes (APC) for each perturbation type.

Queryable 3D mapping. Recent queryable 3D mapping methods [23, 64] typically leverage the results from 2D referring models and fuse them to a 3D map. Thereby, the failure of 2D referring perception can directly obstacle the success of 3D tasks. As shown in Fig. 6, we visualize the state-of-the-art queryable 3D mapping method, ConceptFusion [23]. We notice that with the accurate text query "orange sofa" ConceptFusion can successfully locate the object while when misspells occur in the query, an obvious performance degradation can be observed.

5.3 Perturbation Analysis

Perturbation type. We conduct ablation studies on performance degradation against different perturbation types. Fig. 7 illustrates the aggregated average performance deviation among all benchmarked methods under visual, textual, and acoustic perturbations, respectively. For visual perturbations, we notice the frost effect (FT) induced the most substantial performance degradation, which is likely attributable to its extensive occlusion and distortion effects in the whole image. Conversely, color-based perturbations, such as brightness (BR) and saturate (SA), exhibit marginal impact on performance, possibly due to the preservation of shape information. For textual perturbations, our analysis indicated that all types of perturbations lead to comparable levels of degradation. The character missing (CA) has a slightly greater impact. For acoustic perturbations, we notice that most perturbation types only show a marginal impact on the performance. We consider this is because the target objects in AVS-s4/ms3 are salient objects, allowing models to rely predominantly on visual cues for object localization, even when the acoustic guidance lacks precision. Notably, the impulse response perturbation was associated with a significant performance drop, which could stem from the failure in feature extraction due to the impulse signal. Such abnormal acoustic features could potentially disrupt subsequent multimodal interactions.

Perturbation correlation. To delve deeper into the nature of performance degradation, we computed correlation matrix for each modality, utilizing the Pearson product-moment correlation coefficients calculated from the performance under various perturbation types to populate the matrix. Fig. 8 displays the correlation matrix for textual perturbations.

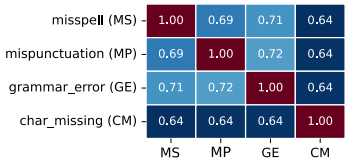


Fig. 8: Correlation of performance of textual perturbations.

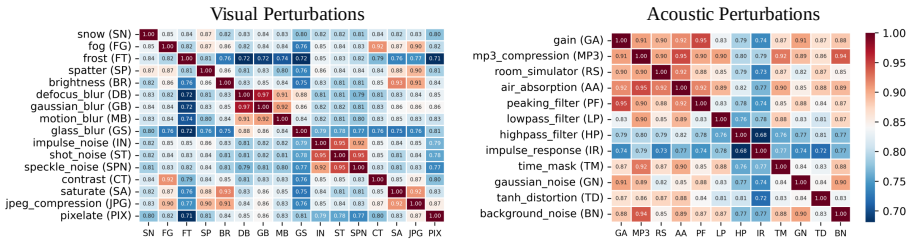


Fig. 9: Correlation matrix of degraded performance with visual and acoustic perturbations. We concatenate the sample-level performance of all methods on all testing datasets as a feature vector to calculate Pearson product-moment correlation.

Despite the average performance changes (in Fig. 7) being similar across different types, the low correlation between the types of performance degradation suggest that each perturbation affects the model in a distinct manner. The correlations for visual perturbations and the ones for acoustic perturbations are presented in Fig. 9 respectively. For visual perturbations, we notice several instances of high correlations. For instance, the group of blurring perturbations—defocus blur (DB), gaussian blur (GB), motion blur (MB), and glass blur (GS)—are highly correlated with one another. For the acoustic perturbations, impulse response (IR) and highpass filter (HF) exhibit different patterns compared to other perturbations, indicating unique effects on performance degradation. We systematically explored the correlations among perturbations and hope the analysis can provide insights for wiser evaluation in the future.

Static v.s. Dynamic. To analyze the dynamic changing of perturbation types in videos, we conduct an ablation study to explore the impact of perturbation changing across frames (for visual perturbation only). As shown in Fig. 10, we calculate the APC across all models on the Ref-DAVIS dataset. We notice the model performance decreases more with dynamic perturbations which can be due to the lack of temporal consistency resulting from various perturbation types.

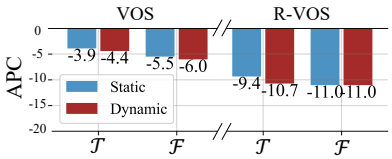


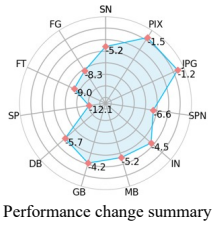
Fig. 10: Average performance change (APC) under static and dynamic perturbations.

5.4 R^2-Agent: Evaluation Assistant

User study of data selection. To evaluate the data selection performance of R^2-Agent, we manually annotate an evaluation set. Since the instruction can be ambiguous in practice, we measure the alignment between R^2-Agent and human by (1) ACC: accuracy between the human-annotated samples and R^2-Agent predictions and (2) Rate: rating the rationality of the prediction

Table 7: Data selection performance.

# Ver. Iter.	✗		1		2		3	
Score	ACC	Rate	ACC	Rate	ACC	Rate	ACC	Rate
Data Sample	0.53	0.78	0.64	0.81	0.77	0.85	0.80	0.91
Perturbation	0.65	0.87	0.77	0.86	0.86	0.91	0.87	0.93
Metric Func.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00



Scenario: Scenario: Car driving scenario with a fast motion
Perturbation: snow (SN), fog (SG), frost (FT), spatter (SP), motion blur (MB),...
Evaluation Metric: mIOU (mean Intersection over Union)

Analysis: IoU Change: The spatter has a largest -12.1 change in mIoU suggesting that the high level of spatter effect has negatively impacted the model's ability. The spatter effect can lead to 1. Boundary Ambiguity...2. Fragmentation...3. Texture change...For jpeg compression and pixelation have a marginal effect...
Case Analysis: The case with caption: "a red car is far away on the left" suffers severe performance drop under...The small size of the car may account for the...
Recommendation: Improve the resilience of models to compression artifacts through enhanced training, including a wider variety of...

Fig. 11: Example of R^2-Agent report. The evaluation setting and metrics are fed to R^2-Agent with a set of hard-coded prompts `<ANA>` to facilitate R^2-Agent generate the human-like report. Beyond the hard-coded analysis, further QA is also supported.

from R^2-Agent (1 if reasonable else 0). As shown in Tab. 7, we ablate on the data verification iteration number. For the metrics function selection, since there are only several hard-coded metrics that can be selected, the accuracy remains perfect while, for data sample and perturbation, we notice a direct performance gain when increasing the iteration number. We notice the performance becomes saturated after 2 iterations, thus we leverage iteration number 2 as the final design choice. More details can be found in the Appendix.

Automatic model analysis. With the evaluation setting and corresponding results, R^2-Agent can generate a detailed evaluation report. We demonstrate an evaluation report example in Fig. 11. To make R^2-Agent think in a human-like manner, we leverage a set of chain-of-thought prompts to instruct R^2-Agent to give reasonable and detailed analysis based on the given evaluation results. We notice that R^2-Agent can understand the evaluation metrics and give a human-like analysis which potentially helps to reduce the cost of model evaluation in practice. Furthermore, given the strong dialogue capability of LLMs, further question-answering about the results and report is also feasible. We demonstrate more qualitative results in the Appendix.

6 Conclusion

In this work, we introduced R^2-Bench, a comprehensive benchmark for evaluating the robustness of visual referring perception models against various perturbations. Our contributions include a detailed taxonomy of perturbations, a customizable perturbation synthesis toolbox, systematic perturbation analysis, and R^2-Agent, a novel evaluation assistant based on large language models. Through extensive experiments, we observed and analyzed the intrinsic characteristics of different perturbations to current models and highlighted the importance of robustness in referring perception tasks. A benchmark with 5 popular referring tasks is also provided to facilitate future research. In addition, the data construction and analysis of robustness evaluation can be further simplified with our proposed R^2-Agent. Our findings underscore the necessity for robustness in the deployment of intelligent systems in real-world scenarios, offering a foundation for future advancements in the field.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 7
2. Ahn, H., Kwon, O., Kim, K., Jeong, J., Jun, H., Lee, H., Lee, D., Oh, S.: Visually grounding language instruction for history-dependent manipulation. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 675–682. IEEE (2022) 2
3. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017) 4
4. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European conference on computer vision (ECCV). pp. 435–451 (2018) 4
5. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4985–4995 (2022) 4, 10
6. Chen, F., Zhang, H., Hu, K., Huang, Y.k., Zhu, C., Savvides, M.: Enhanced training of query-based object detection via selective query recollection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23756–23765 (2023) 2
7. Chen, F., Zhang, H., Li, Z., Dou, J., Mo, S., Chen, H., Zhang, Y., Ahmed, U., Zhu, C., Savvides, M.: Unitail: Detecting, reading, and matching in retail scene (2022), https://arxiv.org/abs/2204.00298 2
8. Chen, F., Zhang, H., Yang, Z., Chen, H., Hu, K., Savvides, M.: Rtgen: Generating region-text pairs for open-vocabulary object detection (2024), https://arxiv.org/abs/2405.19854 2
9. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. arXiv preprint arXiv:2310.12982 (2023) 4, 10, 11
10. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: ICCV (2023) 10
11. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European Conference on Computer Vision. pp. 640–658. Springer (2022) 4, 10
12. Cheng, Y., Wang, R., Pan, Z., Feng, R., Zhang, Y.: Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3884–3892 (2020) 4
13. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) 9, 11
14. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2694–2703 (2023) 2, 4
15. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16321–16330 (2021) 4
16. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023) 3

17. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer. arXiv preprint arXiv:2307.01146 (2023) 2, 11
18. Han, M., Wang, Y., Li, Z., Yao, L., Chang, X., Qiao, Y.: Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13414–13423 (2023) 4
19. Handa, A., Whelan, T., McDonald, J., Davison, A.: A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: IEEE Intl. Conf. on Robotics and Automation, ICRA. Hong Kong, China (May 2014) 9, 11
20. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019) 6
21. Hu, Y., Lin, F., Zhang, T., Yi, L., Gao, Y.: Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. arXiv preprint arXiv:2311.17842 (2023) 2
22. Huang, W., Xia, F., Shah, D., Driess, D., Zeng, A., Lu, Y., Florence, P., Mordatch, I., Levine, S., Hausman, K., et al.: Grounded decoding: Guiding text generation with grounded models for robot control. arXiv preprint arXiv:2303.00855 (2023) 2
23. Jatavallabhula, K.M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Maalouf, A., Li, S., Iyer, G., Saryazdi, S., Keetha, N., et al.: Conceptfusion: Open-set multimodal 3d mapping. arXiv preprint arXiv:2302.07241 (2023) 2, 4, 11, 12
24. Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. In: NeurIPS (2023) 4
25. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: ACCV (2018) 9, 11
26. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 4
27. Li, K., Yang, Z., Chen, L., Yang, Y., Xun, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. arXiv preprint arXiv:2309.09709 (2023) 4, 11
28. Li, M., Li, S., Zhang, X., Zhang, L.: Univs: Unified and universal video segmentation with prompts as queries. arXiv preprint arXiv:2402.18115 (2024) 4
29. Li, X., Cao, H., Zhao, S., Li, J., Zhang, L., Raj, B.: Panoramic video salient object detection with ambisonic audio guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1424–1432 (2023) 2
30. Li, X., Lin, C.C., Chen, Y., Liu, Z., Wang, J., Singh, R., Raj, B.: Paintseg: Painting pixels for training-free segmentation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 2, 4
31. Li, X., Wang, J., Li, X., Lu, Y.: Hybrid instance-aware temporal fusion for online video instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1429–1437 (2022) 2
32. Li, X., Wang, J., Li, X., Lu, Y.: Video instance segmentation by instance flow assembly. IEEE Transactions on Multimedia **25**, 7469–7479 (2022) 2
33. Li, X., Wang, J., Xu, X., Li, X., Raj, B., Lu, Y.: Robust referring video object segmentation with cyclic structural consensus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22236–22245 (2023) 2, 10
34. Li, X., Wang, J., Xu, X., Peng, X., Singh, R., Lu, Y., Raj, B.: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. arXiv preprint arXiv:2310.00132 (2023) 2, 4, 11

35. Li, X., Wang, J., Xu, X., Yang, M., Yang, F., Zhao, Y., Singh, R., Raj, B.: Towards noise-tolerant speech-referring video object segmentation: Bridging speech and text. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2283–2296 (2023) 4, 11
36. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., Shi, S.: Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118 (2023) 3
37. Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18653–18663 (2023) 2, 4, 10
38. Liu, J., Wang, Y., Ju, C., Ma, C., Zhang, Y., Xie, W.: Annotation-free audio-visual segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5604–5614 (2024) 4
39. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 4
40. Liu, S., Hasan, A., Hong, K., Wang, R., Chang, P., Mizrachi, Z., Lin, J., McPherson, D.L., Rogers, W.A., Driggs-Campbell, K.: Dragon: A dialogue-based robot for assistive navigation with visual language grounding. IEEE Robotics and Automation Letters (2024) 2
41. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016) 9, 10
42. Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Spectrum-guided multi-granularity referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 920–930 (2023) 4, 10, 11
43. Mo, S., Morgado, P.: A closer look at weakly-supervised audio-visual source localization. arXiv preprint arXiv:2209.09634 (2022) 4
44. Pan, W., Shi, H., Zhao, Z., Zhu, J., He, X., Pan, Z., Gao, L., Yu, J., Wu, F., Tian, Q.: Wnet: Audio-guided video object segmentation via wavelet-based cross-modal denoising networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1320–1331 (2022) 4
45. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017) 10
46. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4358–4366 (2018) 4
47. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 208–223. Springer (2020) 4, 9
48. Sun, J., Huang, D.A., Lu, B., Liu, Y.H., Zhou, B., Garg, A.: Plate: Visually-grounded planning with transformers in procedural tasks. IEEE Robotics and Automation Letters **7**(2), 4924–4930 (2022) 2
49. Tang, J., Zheng, G., Yang, S.: Temporal collection and distribution for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15466–15476 (2023) 4

50. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 7
51. Tziafas, G., Kasaei, H.: Few-shot visual grounding for natural human-robot interaction. In: 2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). pp. 50–55. IEEE (2021) 2
52. Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X.S., Liang, Y.: Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. Advances in Neural Information Processing Systems **36** (2024) 7
53. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022) 8
54. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2761–2770 (2023) 2, 4, 10, 11
55. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4984 (2022) 4, 10
56. Wu, J., Jiang, Y., Yan, B., Lu, H., Yuan, Z., Luo, P.: Segment every reference object in spatial and temporal spaces. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2538–2550 (2023) 5
57. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023) 7
58. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. arXiv preprint arXiv:2312.00863 (2023) 2, 4
59. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 9, 11
60. Xu, X., Wang, J., Li, X., Lu, Y.: Reliable propagation-correction modulation for video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2946–2954 (2022) 4
61. Xu, X., Wang, J., Ming, X., Lu, Y.: Towards robust video object segmentation with adaptive object calibration. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2709–2718 (2022) 6
62. Xu, X., Zhang, T., Wang, S., Li, X., Chen, Y., Li, Y., Raj, B., Johnson-Roberson, M., Huang, X.: Customizable perturbation synthesis for robust slam benchmarking. arXiv preprint arXiv:2402.08125 (2024) 6
63. Xu, Z., Chen, Z., Zhang, Y., Song, Y., Wan, X., Li, G.: Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17503–17512 (2023) 4, 10
64. Yamazaki, K., Hanyu, T., Vo, K., Pham, T., Tran, M., Doretto, G., Nguyen, A., Le, N.: Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. arXiv preprint arXiv:2310.03923 (2023) 4, 12
65. Yamazaki, K., Truong, S., Vo, K., Kidd, M., Rainwater, C., Luu, K., Le, N.: Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3656–3661. IEEE (2022) 4

66. Yamazaki, K., Vo, K., Truong, Q.S., Raj, B., Le, N.: Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3081–3090 (2023) 4
67. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos (2023) 4
68. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022) 2, 4
69. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022) 10
70. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021) 4, 10
71. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems **35**, 36324–36336 (2022) 4, 10, 11
72. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems **35**, 36324–36336 (2022) 5
73. Yao, J., Wang, X., Ye, L., Liu, W.: Matte anything: Interactive natural image matting with segment anything models. arXiv preprint arXiv:2306.04121 (2023) 4
74. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision. pp. 69–85. Springer (2016) 9, 10
75. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023) 4
76. Zhang, J., Cui, Y., Wu, G., Wang, L.: Joint modeling of feature, correspondence, and a compressed memory for video object segmentation. arXiv preprint arXiv:2308.13505 (2023) 5
77. Zhao, Q., Wang, J., Zhang, Y., Jin, Y., Zhu, K., Chen, H., Xie, X.: Competeai: Understanding the competition behaviors in large language model-based agents. arXiv preprint arXiv:2310.17512 (2023) 3
78. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., et al.: Audio-visual segmentation with semantics. arXiv preprint arXiv:2301.13190 (2023) 4, 9
79. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation. In: European Conference on Computer Vision (2022) 2, 4, 11
80. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8778–8787 (2021), https://api.semanticscholar.org/CorpusID:232093016 2
81. Zhu, C., Chen, F., Shen, Z., Savvides, M.: Soft anchor-point object detection. In: European Conference on Computer Vision (2019), https://api.semanticscholar.org/CorpusID:208512715 2

82. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023) 4, 10
83. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023) 4, 10
84. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024) 4, 10