Towards Motion-aware Referring Image Segmentation

Chaeyun Kim[1,2*] **Seunghoon Yi**[1*] **Yejin Kim**[1] **Yohan Jo**[1] **Joonseok Lee**[1†]

[1]Seoul National University [2]AIM Intelligence

Abstract

Referring Image Segmentation (RIS) requires identifying objects from images based on textual descriptions. We observe that existing methods significantly underperform on motion-related queries compared to appearance-based ones. To address this, we first introduce an efficient data augmentation scheme that extracts motion-centric phrases from original captions, exposing models to more motion expressions without additional annotations. Second, since the same object can be described differently depending on the context, we propose Multimodal Radial Contrastive Learning (MRaCL), performed on fused image-text embeddings rather than unimodal representations. For comprehensive evaluation, we introduce a new test split focusing on motion-centric queries, and introduce a new benchmark called M-Bench, where objects are distinguished primarily by actions. Extensive experiments show our method substantially improves performance on motion-centric queries across multiple RIS models, maintaining competitive results on appearance-based descriptions. Codes are available at https://github.com/snuviplab/MRaCL

1 Introduction

Referring Image Segmentation (RIS) is a task to identify and segment a target object described by a natural language expression from a given image. This task is inherently multimodal, requiring the model to associate visual and linguistic cues in a compositional (*i.e.*, from combined linguistic elements) and contextual (*i.e.*, based on its relation to the scene) manner. With the advancement of vision-language models, the overall performance of RIS models has shown notable improvement (Yang et al., 2022; Wang et al., 2022; Hu et al., 2023; Ha et al., 2024; Yue et al., 2024).

A key challenge in RIS stems from the complexity and variety of linguistic cues used in referring expressions. In other words, the same object can be referred to in vastly different ways, including visual appearance (*red*, *smiling*), multi-object relations (*on the left, in front of*), and motions (*walking, jumping*). For instance, a person might be referred to as "the person wearing an orange jacket" (appearance-centric) or "the person bending over and touching shoes" (motion-centric).

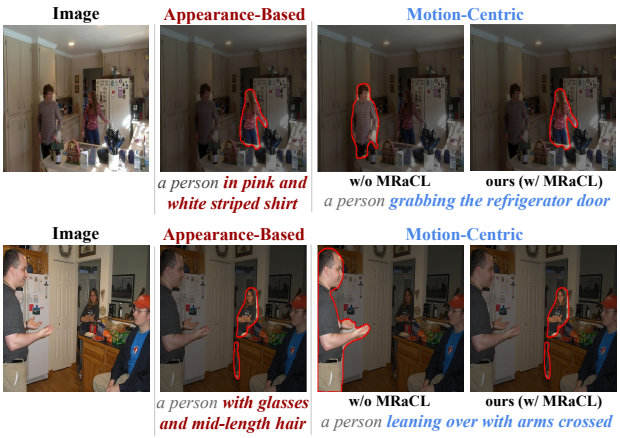


Figure 1: **Appearance-based (left) vs. motion-centric (right) queries in RIS.** Existing methods handle the former well but struggle with the latter.

	mIoU		oIoU	
Models	**Static**	**Motion**	**Static**	**Motion**
CRIS (Wang et al., 2022)	68.66	52.53	65.67	45.98
LAVT (Yang et al., 2022)	71.52	53.28	70.84	46.58
DMMI (Hu et al., 2023)	73.41	51.62	71.54	45.83
ASDA (Yue et al., 2024)	76.26	54.08	75.19	48.03

Table 1: Performance gap between appearance-centric and motion-centric queries on G-Ref UMD test set.

*Equal contribution.

†Correspond to: joonseok@snu.ac.kr

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

Although both refer to the same target, the type of perception and reasoning required to distinguish this object from others can be significantly different.

Despite recent advances, we observe that even the state-of-the-art RIS models often struggle when motion-centric cues are given. As illustrated in Fig. 1, the model successfully segments the target when described with visual attributes in the second column, while fails when the same target is described by their motion, in the third column. To verify whether this is a general trend, we manually curated two test sets, 300 appearance-centric and 300 motion-centric queries, from the G-Ref (Mao et al., 2016) UMD split. Tab. 1 confirms that all evaluated models show significant performance drops on motion-based queries, revealing an over-reliance on static visual attributes.

Why do the RIS models particularly suffer from motion-centric queries? We hypothesize that insufficient exposure to diverse motion expressions during training is the primary cause. Motion-related descriptions typically appear as verb phrases, and our analysis reveals that motion-oriented expressions have been underrepresented. For instance, RefCOCO and RefCOCO+ contain only 10.4% and 13.8% of motion-centric queries, respectively. G-Ref or Ref-ZOM contain more (40.3% and 38.2% each), but such motion-oriented descriptions are often shadowed by easily identifiable accompanying nouns or subject descriptions. This leaves motion-based expressions underutilized and hinders their ability to learn robust associations with the target.

To address this, we propose a data-augmentation approach to provide richer motion-centric training queries without requiring additional annotation. Inspired by dropout (Srivastava et al., 2014), we propose to extract motion-based sub-part of the original textual description and utilize it as a positive during training, to encourage the model to learn diverse linguistic structures. For instance, from a full expression "a woman in orange shirt bending over", we extract "bending over" as a supplementary positive description, guiding the model to localize targets with motion-oriented expressions without relying on other cues.

However, simply adding motion-centric expressions is insufficient. In RIS, different expressions may refer to the same object only under specific image context. For instance, "a man with blue shirt" and "a man running on the street" may refer to the same individual when conditioned on the image, but these convey different semantics in general. Thus, simply placing embeddings of these two expressions closely is valid only under the given visual context; otherwise, such alignment may lead to incorrect associations.

Therefore, a naive use of the standard contrastive learning to unimodal representations would fail to capture this context-conditioned semantic overlap. To this end, we propose to perform contrastive learning *after fusing* the visual and textual embeddings. Additionally, we observe that the conventional cosine similarity-based objective suffers from gradient saturation, hindering fine-grained training. To resolve this problem, we design an alternative loss directly based on the angle between two embeddings, namely, Multimodal Radial Contrastive Loss (MRaCL).

Combining these ideas, our proposed approach augments with pairs of ⟨image, motion-centric phrase⟩, taking them as positives, and performs contrastive learning on their multimodal embeddings using our novel MRaCL loss. It consistently improves performance across multiple baselines, particularly for motion-centric queries, while maintaining competitive accuracy on static expressions. We further introduce M-Bench, a curated benchmark composed of action-centric RIS examples from video datasets, where the target object must be localized primarily via motion.

We summarize our contributions as follows:

1. We identify the challenge of motion-centric queries in RIS and propose a novel textual augmentation strategy to mitigate underrepresented motion expressions.
2. To better reflect the context-dependent nature of RIS, we propose MRaCL, a novel contrastive loss for image-conditioned semantic alignment of multimodal embeddings.
3. We verify the effectiveness of our method across multiple RIS baselines and datasets, including our newly proposed motion-focused evaluation benchmark, M-Bench.

2 Related Work

Referring Image Segmentation. The early architectures for reference image segmentation (RIS) extracted visual and linguistic features separately using CNNs (Hu et al., 2016; Liu et al., 2017a) and RNNs (Li et al., 2018; Liu et al., 2017a; Margffoy-Tuay et al., 2018; Shi et al., 2018), which were then concatenated to form multimodal features. More recent work has shifted towards Transformer-based backbones (Jing et al., 2021; Liu et al., 2021; Devlin et al., 2019; Kamath et al., 2021) and multiscale features (Chen et al., 2019; Hu et al., 2020; Hui et al., 2020; Ye et al., 2019) to capture finer object details and boundaries.

A central challenge in RIS is the effective fusion of multimodal features. Fusion strategies have progressed from simple concatenation (Hu et al., 2016) to more so-

phisticated techniques. These can be broadly classified into three categories: (1) *Early fusion* methods such as LAVT (Yang et al., 2022) and DMMI (Hu et al., 2023) perform cross-modal interaction within the stages of the vision encoder. (2) *Late fusion* approaches including VLT (Ding et al., 2021), CRIS (Wang et al., 2022), and ReSTR (Kim et al., 2022) encode modalities independently before merging them in a cross-modal decoder. (3) *Multi-level fusion*, employed by CrossVLT (Cho et al., 2023) and CoupAlign (Zhang et al., 2022c) introduces cross-modal interactions at multiple stages of the encoders.

Recent works explore diverse set of architectural combinations. CGFormer (Tang et al., 2023), ASDA (Yue et al., 2024), and VATEX (Nguyen-Truong et al., 2025) organize visual features into language-conditioned tokens to leverage global and local information. ReLA (Liu et al., 2023a) and DMMI (Hu et al., 2023) extend the task to support an arbitrary number of targets. Most recently, DETRIS (Huang et al., 2025) has adopted parameter-efficient fine-tuning on pre-trained encoders with multiscale feature alignment.

Deep Metric Learning. Deep Metric Learning (DML) aims to learn embedding functions that map semantically similar instances closer together while pushing dissimilar ones apart in the feature space. Foundational works such as contrastive loss (Chopra et al., 2005), N-pair loss (Sohn, 2016), and triplet margin loss (Lee et al., 2018, 2020; Ma et al., 2021) established early distance-based objectives using L_p norms or dot-product similarity.

Subsequent approaches adopted softmax-inspired objectives to improve discriminability. For example, Center Loss (Wen et al., 2016), SphereFace (Liu et al., 2017b), and CosFace (Wang et al., 2018) have been widely used in face recognition. Building upon this, ArcFace (Deng et al., 2019) introduced an angular-margin loss to enforce larger angular separations. Recently, AoE (Li and Li, 2024) applied these ideas to semantic textual similarity tasks while addressing gradient saturation issue common in cosine-based losses.

Although the principles of DML have been increasingly applied to multimodal domains, a key limitation remains: standard contrastive objectives, often relying on cosine similarity, suffer from gradient saturation and is thus suboptimal in the inherently anisotropic spaces of pre-trained encoders (Gao et al., 2021; Li and Li, 2024). Our proposed Multimodal Radial Contrastive Loss (MRaCL) addresses this limitation by employing an angle-based similarity metric within a contrastive framework on fused multimodal embeddings, enabling more effective fine-grained alignment for the RIS task, as detailed in Sec. 3.3.

3 Method

Our approach enhances RIS models' understanding of action-centric expressions through two complementary strategies. First, we propose a data augmentation strategy that extracts action-focused phrases from referring expressions. Second, recognizing that semantically different phrases can refer to the same object only within a specific image context, we introduce Multimodal Radial Contrastive Loss (MRaCL) that performs contrastive learning on fused image-text embeddings rather than unimodal representations. Observing that simply exposing the model to more verb phrases is not sufficient, however, we further propose a more direct enhancement in RIS modeling and training. Considering the unique characteristics of RIS that semantically different phrases can refer to the same object conditioned on an image, it is often misleading to simply align textual embeddings referring to the same object. Instead, we propose to align them in conjunction with the conditioned image, so that the model can learn semantic proximity between multimodal embeddings, rather than unimodal textual embeddings. Furthermore, we introduce a novel objective called Multimodal Radial Contrastive Loss (MRaCL), designed to expedite the fine-grained alignment in a shared multimodal embedding space. Together, these strategies enable the model to capture nuanced action semantics and foster more robust compositional understanding in RIS. We describe each idea in detail in this section.

3.1 Data Augmentation with Motion-centric Verb Phrase Retrieval

Existing RIS models often overlook action-centric expressions, as objects can be often distinguished by appearance alone. We hypothesize that exposing a model to more training examples where the target is disambiguated only through its motion will improve the model's motion understanding capabilities. However, this approach can introduce ambiguity when multiple objects of the same category perform similar actions (*e.g.*, two people 'jumping over'). To mitigate this potential issue, we apply a pre-filtering strategy to our training data, retaining only those samples where the target's category appears a single time. The effectiveness of this strategy is validated in Tab. 7. To this end, we propose to augment the training set by adding *verb phrases* which highlights the motion extracted from the original expression.

More formally, given a training example (I, T) where I is an image and T its textual description, we extract verb phrases from T that capture the referent's motion, using an LLM (*e.g.*, LLaMA 3.1-70B (Grattafiori et al., 2024)) with a customized prompt that guides

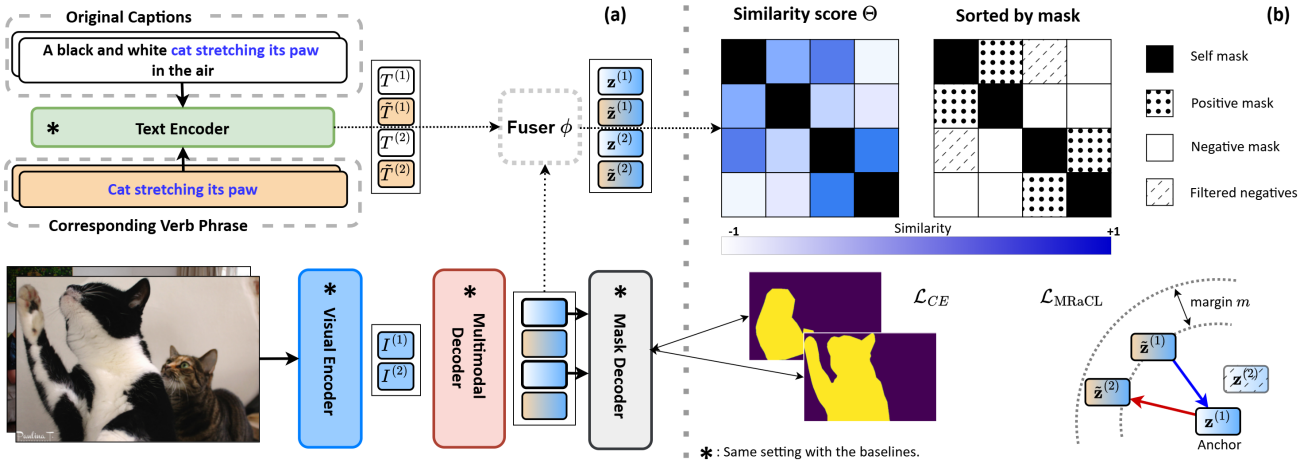


Figure 2: **Overview of the MRaCL framework**, (a) We apply CE Loss to output masks of the original pairs, same as the baselines. The *Fuser* mix text embeddings and cross-modal embeddings from the decoder, returns multimodal representations $\mathbf{z}^{(i)}$. (b) With those multimodal latents, we calculate similarity scores and filter out false negatives. For example, with $\mathbf{z}^{(1)}$ as the anchor, $\mathbf{z}^{(2)}$ is considered as false negative sample and masked out as shown in the bottom right diagram. Here, components with marked as '$*$' indicates that it originated form the baseline model.

it to *retrieve* motion-oriented verb phrases from the original caption, ensuring that they are grounded in the input description. The LLM identifies whether T contains such verb phrase and, if so, extracts the most detailed and contextually relevant span, denoted by $\tilde{T}$. The resulting $\tilde{T}$ may include a single verb ('running'), a verb-object pair ('doing tricks'), or a longer construct with prepositional phrases ('placed on the table'), covering both active motions and passive states.

Crucially, we utilize $(I, \tilde{T})$ as a *supplementary*, not as a replacement for the training example, alongside the original (I, T), thereby *retaining original cues* from T while reinforcing the model's attention to motion-centric semantics through $\tilde{T}$. Training with the supplementary examples, the model is guided to learn that both T and $\tilde{T}$ refer to the same target. By incorporating these partial, motion-centric expressions into training, we encourage the model to correctly localize the target even in the absence of static attributes or complete contextual details.

3.2 Multi-modal Contrastive Learning

We empirically observe, however, that simply providing additional training examples does not help. In fact, naively performing contrastive learning on these augmented data actually drops the RIS performance. This stems from a unique characteristic of RIS: an object can be described with various aspects (*e.g.*, appearance, motion, relation with others), and these expressions are semantically aligned only when they are conditioned on the image. For example, 'a lady wearing yellow shirt' and 'a lady running on the street' convey completely uncorrelated semantics in general, but they may refer to the same object conditioned on a particular image. For this reason, simply learning to align them without grounding in a visual context would confuse the model. We resolve this challenge by implementing a cross-modal contrastive learning framework, which we briefly review first and develop our ideas.

Review of Contrastive Learning. As a representative metric learning method, contrastive learning has become a cornerstone for representation learning. SimCSE (Gao et al., 2021) introduced a effective way to learn sentence representations, subsequently adopted by many variants (Zhou et al., 2022; Zhang et al., 2022a; Chuang et al., 2022; Zhang et al., 2022b; Liu et al., 2023b). Given a batch $\mathcal{B} = \{\mathbf{z}_i\}_{i=1}^n$ of text embeddings, SimCSE creates positive pairs by encoding each caption twice with dropout-based augmentation. Taking the other augmented one from the same text as a positive and all other pairs in the batch as negatives (Chen et al., 2017), SimCSE minimizes the NT-Xent loss (Chen et al., 2020)

$$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{e^{\text{sim}(\mathbf{z}^{(i)}, \tilde{\mathbf{z}}^{(i)})/\tau}}{\sum_{j=1}^{n} e^{\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})/\tau}}, \qquad (1)$$

where $\mathbf{z}^{(i)}$ is the anchor, $\tilde{\mathbf{z}}^{(i)}$ is the positive sample of $\mathbf{z}^{(i)}$, and $\mathbf{z}^{(j)}$ $(j \neq i)$ are negative samples, and τ is the temperature hyperparameter. The $\text{sim}(\mathbf{x}, \mathbf{y})$ is a similarity function between $\mathbf{x}$ and $\mathbf{y}$, *e.g.*, the cosine similarity $\text{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y} / \|\mathbf{x}\| \|\mathbf{y}\|$.

Extension to Multimodal Embedding Space. To ensure that semantically different texts referring to the same object are aligned only within their specific visual context, we propose to conduct contrastive learn-

ing on *fused cross-modal representations* rather than unimodal representations in RIS.

Given an image-text pair (I, T), we extract a cross-modal embedding $\mathbf{z} = \phi(\mathbf{x}_I, \mathbf{x}_T)$, where $\mathbf{x}_I$ and $\mathbf{x}_T$ indicate the image and text features from corresponding pre-trained models (see Sec. 4.1 for details). ϕ is a visual-text fuser, which we use a single projection layer, in Fig. 2(a). $\mathbf{z}$ serves as our anchor, as shown in the lower right of Fig. 2(b). Similarly, we also extract the cross-modal embedding from the augmented pair $(I, \tilde{T})$, yielding a positive pair embedding $\tilde{\mathbf{z}} = \phi(\mathbf{x}_I, \mathbf{x}_{\tilde{T}})$. Other pair embeddings such as $\phi(\mathbf{x}_I^{(i)}, \mathbf{x}_T^{(j)})$ for $j \neq i$ are considered as *negatives*, where i, j are indices within a minibatch. With this redefined anchor, positive and negatives, we can apply contrastive learning, *e.g.*, via minimizing the NT-Xent loss in Eq. (1) or its variants, in the multimodal space.

False Negative Elimination. In the RIS setting, where image–text pairs are usually not in a one-to-one correspondence, semantically similar descriptions often appear across different samples. For instance, "person walking" and "moving on the street" describe similar actions, and they may refer to the similar target from different images. Thus, treating them as a negative may not be desirable.

Inspired by the false negative cancellation strategy (Huynh et al., 2022), we design an optional filtering step to identify and filter out potential false negatives leveraging the model's own multimodal embeddings. Specifically, we first calculate similarity score Θ of anchor embeddings within each mini-batch, as in Fig. 2(b). When the score between an anchor $\mathbf{z}^{(i)}$ and a nominal negative exceeds a threshold ν (empirically 0.5), we consider that pair as a potential *false negative* and exclude from the contrastive loss computation.

3.3 Multimodal Radial Contrastive Learning

Although we have addressed the issue with action-centric queries and contrastive learning from the RIS perspective, the contrastive objective based on cosine similarity remains suboptimal, for two reasons. First, the cosine similarity suffers from *similarity saturation* (Li and Li, 2024). As the similarity approaches ± 1, its gradient significantly diminishes, providing negligible update signals when embeddings are already close (positive pairs) or distant (negative pairs). This hinders learning of fine-grained distinctions crucial for differentiating subtle action variations. Second, multimodal embeddings exhibit *anisotropy*, which tend to cluster in confined regions of the embedding space. Primarily observed in pre-trained language models (Gao et al., 2019; Ethayarajh, 2019; Wang et al., 2019; Li et al., 2020), it also occurs in RIS models. For example, empirical pairwise angular distances between embeddings from CRIS (in Fig. 3(a)), reveals that most embeddings are tightly clustered within a small angular distance. This exacerbates the saturation problem, making it particularly challenging to differentiate fine-grained action expressions.

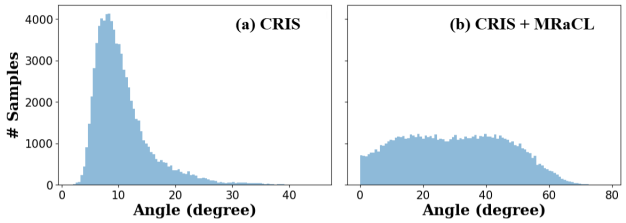


Figure 3: **Anisotropy phenomenon observed in a baseline model (CRIS)**. We plot the distribution of pairwise angular distances of 100K pairs, (a) from the original CRIS and (b) with our MRaCL loss. We clearly see that the model utilizes a significantly wider representation space using our MRaCL loss.

To address these issue, we adopt the *angular distance* as our similarity metric, building upon AoE (Li and Li, 2024). Specifically, for two normalized multimodal representations, $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$, we define our similarity metric $\theta_{i,j}$ as

$$\theta_{i,j} = \tfrac{\pi}{2} - \angle(\mathbf{z}^{(i)}, \mathbf{z}^{(j)}) = \tfrac{\pi}{2} - \cos^{-1}\big(\mathrm{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})\big), \quad (2)$$

where $\angle(\mathbf{x}, \mathbf{y})$ indicates the angle between the two vectors $\mathbf{x}$ and $\mathbf{y}$, and $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. We subtract from $\pi/2$ to convert the angle to a similarity score. This choice mitigates the similarity saturation problem, as the derivative $d\theta_{i,j}/d\mathbf{z}$ does not vanish unless $\theta_{i,j} = \pi/2$, thereby maintaining substantial gradients even for closely aligned embeddings.

For a more robust alignment, we introduce a margin $m \geq 0$ on the positive pair similarities $\theta_{i,\tilde{i}}$. This enforces a minimum angular separation between positive and negative pairs, encouraging the model to learn more discriminative representations. Also, it mitigates anisotropy by explicitly penalizing embeddings that are too close, driving them to utilize a broader embedding space.

Replacing the similarity score with Eq. (2) and combining all the elements, we get our Multimodal Radial Contrastive Loss (MRaCL):

$$\mathcal{L}_{\mathrm{MRaCL}} = -\log \frac{\exp\big((\theta_{i,\tilde{i}} - m)/\tau\big)}{\exp\big((\theta_{i,\tilde{i}} - m)/\tau\big) + \sum_{j \neq i} \exp(\theta_{i,j}/\tau)}, \quad (3)$$

where $\theta_{i,\tilde{i}} = \pi/2 - \angle(\mathbf{z}^{(i)}, \tilde{\mathbf{z}}^{(i)})$ following Eq. (2).

Finally, integrating with the conventional segmentation loss $\mathcal{L}_{\mathrm{seg}}$ (*e.g.*, pixel-level mask cross-entropy) completes our full training objective: $\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \alpha \cdot \mathcal{L}_{\mathrm{MRaCL}}$, with a balancing weight α.

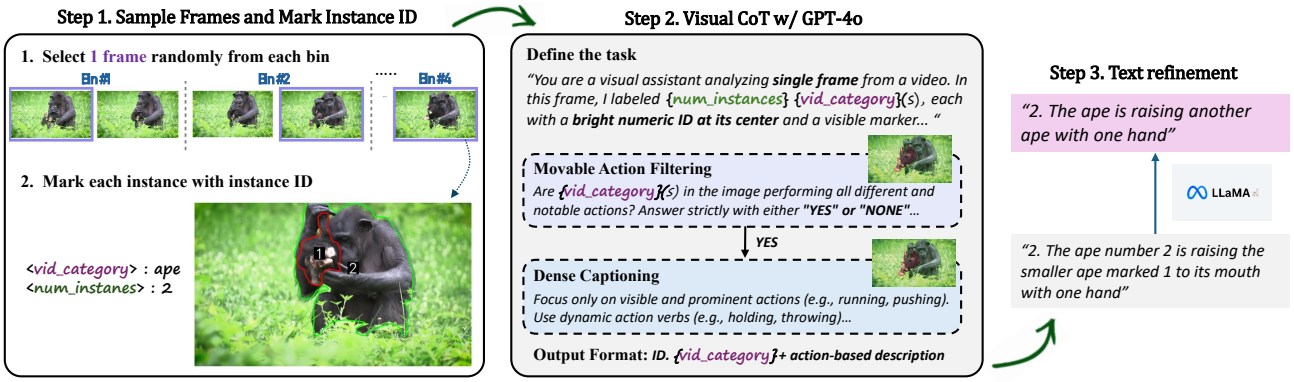


Figure 4: **Overview of the MBench dataset annotation pipeline.**

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate our method on four widely used RIS benchmarks: RefCOCO (Yu et al., 2016), RefCOCO+ (Kazemzadeh et al., 2014), G-Ref (Mao et al., 2016) and Ref-ZOM (Hu et al., 2023). RefCOCO contains positional phrases like "front" or "third from the right," while RefCOCO+ excludes such cues. G-Ref features longer and more descriptive expressions (9–10 words on average), and Ref-ZOM allows cases with no target or multiple targets. Among these, G-Ref and Ref-ZOM include more motion-centric cues and object relationships, making them particularly suitable for evaluating our approach.

Motion-Oriented Benchmarks. To systematically evaluate motion understanding in RIS, we introduce two complementary benchmarks. First, **M-Ref** consists of 325 appearance-centric and 325 motion-centric queries manually sampled from the G-Ref (UMD) test set, enabling controlled comparison between query types (examples in App. A).

Second, we construct **M-Bench**, a comprehensive motion-centric benchmark built upon three video object segmentation datasets: A2D-Sentences (Gavrilyuk et al., 2018), Ref-Youtube-VOS (Seo et al., 2020), and ViCaS (Athar et al., 2024). Leveraging their pixel-level mask annotations, we generate fine-grained action-oriented captions aligned with each object through a two-stage pipeline (Fig. 4).

We initially sample representative frames and retain only those containing at least two instances of the same movable category (*e.g.*, multiple persons), ensuring that targets cannot be identified by category alone. Then, we employ visual chain-of-thought prompting with an MLLM to produce a concise, motion-centric description for each instance, followed by refinement with a lightweight language model. Every description is manually verified, and we discard samples whose captions do not reflect the actual behavior or where multiple instances exhibit indistinguishable actions. This yields 8,200 evaluation-only samples designed to assess action-based expression understanding. Full details are in Apps. B and C.

Evaluation Metrics. We adopt three standard metrics for RIS: mIoU, oIoU, and Prec@p. The mIoU is the average IoU values computed per image, while oIoU measures the ratio of total intersection to total union across all test objects (samples). Since oIoU tends to favor larger objects, we use mIoU as a complementary size-balanced measure. Prec@p with $p \in \{0.5, 0.7, 0.9\}$ indicates the percentage of samples with IoU above threshold p.

Implementation Details. We evaluated five state-of-the-art baselines with and without our method: LAVT (Yang et al., 2022), CRIS (Wang et al., 2022), DMMI (Hu et al., 2023), ASDA (Yue et al., 2024), and DETRIS (Huang et al., 2025). These models represent different fusion strategies: LAVT and DMMI employ early fusion, CRIS and ASDA utilize late fusion, while DETRIS adopts multi-level fusion.

We keep original architectures and hyperparameters intact, adding only a single linear projection layer to produce multimodal embeddings for our MRaCL objective. For MRaCL-specific hyperparameters, we set $\alpha = 0.1$, margin $m = 12$, and temperature $\tau = 0.07$ across all datasets, empirically determined through ablation in Sec. 4.3. All baselines are reproduced on NVIDIA RTX A6000 GPUs for fair comparison. Additional results on the recent LatentVG (Yu et al., 2025) are provided in App. D.

4.2 Results and Analysis

Overall Comparison on Existing Benchmarks. The four leftmost column groups in Tab. 2 compares the average oIoU of competing RIS models with and

Method	+MRaCL	RefCOCO(UNC) oIoU val	RefCOCO(UNC) oIoU testA	RefCOCO(UNC) oIoU testB	RefCOCO+(UNC+) oIoU val	RefCOCO+(UNC+) oIoU testA	RefCOCO+(UNC+) oIoU testB	G-Ref(UMD) oIoU val(U)	G-Ref(UMD) oIoU test(U)	Ref-ZOM oIoU test(F)	M-Ref mIoU static	M-Ref mIoU motion	M-Ref oIoU static	M-Ref oIoU motion	MBench mIoU	MBench oIoU
CRIS	✗	66.68	70.62	59.93	56.94	64.20	46.97	55.91	57.05	58.62	68.66	52.53	65.67	45.98	46.21	47.62
(Wang et al., 2022)	✓	67.61	71.89	61.81	57.26	65.04	47.25	58.05	58.93	60.97	70.14	54.25	67.13	49.05	47.25	49.43
	Diff	**+0.93**	**+1.27**	**+1.88**	**+0.32**	**+0.84**	**+0.28**	**+2.14**	**+1.88**	**+2.35**	**+1.48**	**+1.94**	**+1.46**	**+3.07**	**+1.04**	**+1.81**
LAVT	✗	72.72	75.74	68.56	61.86	67.97	54.30	61.24	62.09	64.48	71.52	53.28	70.84	46.58	52.28	57.48
(Yang et al., 2022)	✓	73.07	76.13	69.41	63.05	67.95	54.36	63.93	64.12	66.24	72.85	56.21	71.41	50.11	53.32	59.17
	Diff	**+0.35**	**+0.39**	**+0.85**	**+1.19**	**+0.28**	**+0.06**	**+2.69**	**+2.03**	**+1.76**	**+1.33**	**+2.93**	**+0.57**	**+3.53**	**+1.04**	**+1.69**
DMMI	✗	72.85	75.53	69.53	63.18	67.98	55.53	61.75	62.54	66.94	73.41	51.62	71.54	45.83	52.89	55.94
(Hu et al., 2023)	✓	73.31	76.26	69.96	63.51	68.66	56.15	63.24	64.98	68.26	74.68	53.72	72.73	47.42	54.16	57.45
	Diff	**+0.46**	**+0.73**	**+0.43**	**+0.33**	**+0.68**	**+0.62**	**+1.49**	**+2.44**	**+1.32**	**+1.27**	**+2.10**	**+1.19**	**+1.59**	**+1.27**	**+1.51**
ASDA	✗	75.08	77.58	71.21	66.97	71.81	57.92	65.71	65.38	57.94	76.26	54.08	75.19	48.03	52.92	37.29
(Yue et al., 2024)	✓	75.82	78.15	72.17	67.98	72.89	58.79	66.98	67.26	58.64	76.79	58.95	76.25	53.15	54.17	38.85
	Diff	**+0.74**	**+0.57**	**+0.96**	**+1.01**	**+1.08**	**+0.87**	**+1.27**	**+1.88**	**+0.70**	**+0.53**	**+4.07**	**+1.06**	**+5.12**	**+1.25**	**+1.56**
DETRIS	✗	74.06	77.15	71.08	64.62	71.86	56.35	64.29	64.98	67.36	75.87	62.01	75.13	56.27	58.36	61.22
(Huang et al., 2025)	✓	74.94	77.83	71.90	65.25	72.09	56.88	64.94	65.61	68.35	77.53	65.23	76.42	58.43	59.69	63.03
	Diff	**+0.88**	**+0.68**	**+0.82**	**+0.63**	**+0.23**	**+0.53**	**+0.65**	**+0.63**	**+0.99**	**+1.66**	**+3.22**	**+1.29**	**+2.16**	**+1.33**	**+1.81**

Table 2: **Overall comparison of the RIS models**. Green cells indicate statistically significant improvement, and gray cells mean neutral. (Red cells mean significant performance drop, but we have no such case.)

without MRaCL loss. For all datasets and models, MRaCL consistently improves performance, demonstrating its general effectiveness.

An interesting observation is that the degree of improvement correlates with the ratio of motion-centric queries per dataset. On G-Ref and Ref-ZOM, which are relatively rich in verb-centric phrases, our method achieves substantial gains, boosting the baseline models by 1.71 and 1.42 oIoU points on average. Notably, this trend holds even on verb-sparse datasets; on RefCOCO and RefCOCO+, adding MRaCL loss still provides consistent improvements with an average gain of 0.78 and 0.59 oIoU points, respectively.

This implies models trained with MRaCL better capture the nuances of verb-driven relationships and their visual grounding. Also, our method enhances general multi-modal alignment beyond motion understanding, benefiting RIS performance across diverse query types.

Evaluation on Action-centric Benchmarks. The rightmost 3 columns in Tab. 2 compares the RIS performance on our motion-centric benchmarks: M-Ref and M-Bench. Results on M-Ref demonstrate that our method improves performance on both static and motion splits, with substantially larger gains on motion-centric queries. This confirms that our approach is particularly effective when objects must be disambiguated through motion-related cues rather than static attributes describing their appearance.

Lastly, recall that MBench features highly ambiguous scenarios, with multiple similar objects performing diverse actions. The evaluation results on MBench in the rightmost column of Tab. 2 also show consistent improvements across all models with our method. Even in this challenging scenario, our method leads to significant gains, showing its effectiveness in helping RIS models focus on motion expressions.

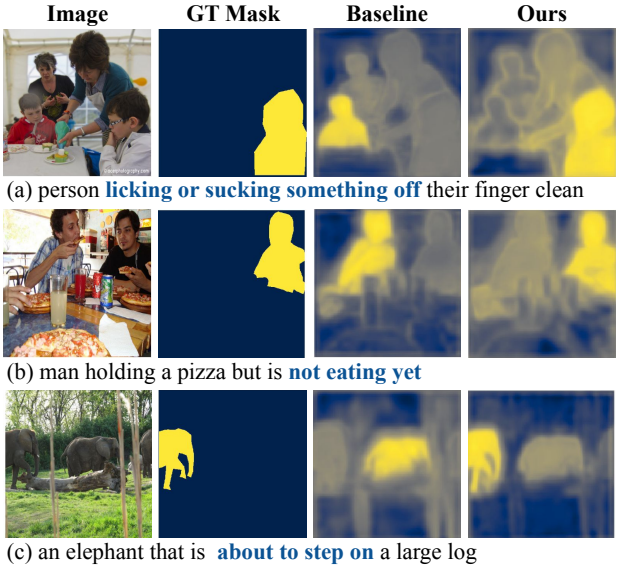


Figure 5: Visualization of activation maps with and without MRaCL on CRIS.

Qualitative Analysis. Figs. 5 and 6 both illustrate the effect of our method. We observe two consistent patterns of improvement. First, our method strengthens ***action-level disambiguation***. As seen in Fig. 6(a,c), CRIS with MRaCL correctly segments the target among multiple instances of the same category by grounding fine-grained action cues, where the baseline conflates visually similar subjects. The activation maps in Fig. 5(a) further confirm this, as our method produces sharply focused activations on the acting person rather than diffusing across co-occurring instances. Second, MRaCL better captures ***nuanced motion expressions*** that go beyond simple verb matching. In Fig. 5(b,c), queries involv-

Method	mIoU	oIoU	Prec@ 0.5	Prec@ 0.7	Prec@ 0.9
CRIS	59.17	55.91	67.95	55.45	15.27
CRIS + L2	58.99	55.17	67.85	55.09	14.87
CRIS + SimCSE	59.77	56.09	68.15	55.45	14.42
CRIS + InfoNCE (MCC)	59.46	55.75	67.87	55.35	14.95
CRIS + MRaCL (ours)	**60.87**	**58.05**	**69.73**	**57.45**	**15.85**

Table 3: **Comparison with various contrastive objectives**, measured on G-Ref validation set (UMD).

Figure 6: Qualitative results from MBench test set.

ing negation ("not eating yet") and prospective intent ("about to step on") require a deeper understanding of the described action context. Our method yields well-delineated activations aligned with the correct target in both cases, whereas the baseline fails to distinguish these subtle cues. See App. E for additional examples.

Comparison with other Contrastive Losses. In Tab. 3, we compare various contrastive loss objectives on the G-Ref UMD validation set. Starting from vanilla CRIS trained with cross-entropy only, we evaluate three contrastive alternatives: L2-based (Euclidean), SimCSE (cosine-based) (Gao et al., 2021), and InfoNCE-based Meaning Consistency Constraint (MCC) (Oord et al., 2018; Nguyen-Truong et al., 2025), alongside our MRaCL.

While L2-based loss slightly degrades performance and SimCSE and InfoNCE (MCC) yield marginal gains, MRaCL substantially outperforms all alternatives across all metrics. We attribute this to the following factors: 1) distance-based losses such as L2 allow flexible representation scaling but lack the angular sensitivity needed for fine-grained multimodal alignment; and 2) cosine similarity, used in both SimCSE and MCC, is prone to gradient saturation and anisotropic clustering (Chen et al., 2024), as further discussed in Sec. 3.2. In contrast, MRaCL leverages angular distance with explicit margin penalties, producing more discriminative embeddings that better capture subtle multimodal semantics.

Ratio	mIoU	oIoU	P@0.5	P@0.7	P@0.9
0	**60.87**	**58.05**	**69.73**	**57.45**	**15.85**
0.03	59.22	56.51	67.95	55.03	14.89
0.1	46.84	42.42	51.35	44.76	10.03
0.2	43.94	41.25	47.39	42.08	10.34

Table 4: **Ablation on semi-hard negative mining.** A ratio of 0 corresponds to our default augmentation-based alignment without semi-hard negatives.

Augment	MRaCL	Filtering	mIoU (%)	oIoU (%)
✗	✗	✗	59.28	55.91
✗	✓	✗	59.45	56.02
✓	✗	✗	59.31	57.34
✗	✓	✓	59.89	56.11
✓	✓	✗	60.18	57.33
✓	✓	✓	**60.87**	**58.05**

Table 5: **Ablation on proposed components on G-Ref.** Motion-centric augmented phrases (Sec. 3.1), angular distance-based MRaCL loss (Sec. 3.3), and filtering for potential false negatives (Sec. 3.2).

Positive Alignment vs. Negative Mining. We examine whether semi-hard negative mining could further improve performance, considering negatives of two types: a different object performing the same action, or the same object performing a contradictory action. To approximate these, we replace the verb in each ground-truth caption with a randomly selected one, and train CRIS (Wang et al., 2022) on G-Ref (UMD) with varying mixing ratios of such negatives.

As shown in Sec. 4.2, incorporating semi-hard negatives consistently degrades performance, even at a ratio as small as 0.03. We attribute this to the current state of RIS models: hard negative learning is generally effective when a model already performs coarsely and needs finer discrimination. However, as established in Sec. 1 (Tab. 1), existing RIS models largely ignore motion cues and rely predominantly on static appearance. Since these models have not yet learned to leverage motion semantics at all, hard negatives can act as noise rather than useful training signal. This motivates our positive alignment strategy, which first provides a clear directional signal toward motion understanding as a necessary foundation before any discriminative negative objective can be effective.

4.3 Ablation Study

Effectiveness of Components. We analyze the contribution of each component in our framework using CRIS (Wang et al., 2022) on G-Ref (UMD) validation set. Specifically, we compare combinations of three design factors: 1) using action-centric phrases as additional contrastive anchors (Augmentation; Sec. 3.1), 2) applying our MRaCL loss based on angular distance

Hyperparameter		val(UMD)		Prec(val)		
		mIoU	oIoU	0.5	0.7	0.9
Margin (m)	4	59.13	56.04	68.05	54.82	15.01
	8	59.66	56.80	68.71	56.85	15.08
	10	60.08	57.51	69.36	57.16	15.12
	12	**60.87**	**58.05**	**69.73**	**57.45**	**15.85**
	15	60.06	56.83	69.04	57.09	15.22
	20	59.26	55.88	67.97	55.56	15.24
Filtering Threshold (ν)	0.35	59.54	55.93	68.16	55.29	15.51
	0.40	60.04	56.58	69.02	56.48	15.57
	0.45	60.11	57.02	69.28	56.84	15.42
	0.50	**60.87**	**58.05**	**69.73**	**57.45**	**15.85**
	0.60	58.68	55.02	67.91	55.33	14.85
	0.70	59.52	56.12	68.01	56.35	15.62
Metric Loss Weight (α)	0.05	60.17	57.25	69.12	56.33	15.93
	0.10	**60.87**	**58.05**	**69.73**	**57.45**	**15.85**
	0.15	60.13	57.68	68.63	57.38	15.53
	0.10	59.64	56.13	68.51	55.47	15.23
	0.25	59.43	55.54	68.35	55.13	14.01

Table 6: **Ablation on hyperparameter selection.** Experiments were conducted on G-Ref.

(MRaCL; Sec. 3.3), and 3) similarity-based filtering to eliminate potential false negatives (Filtering; Sec. 3.2).

In Tab. 5, applying MRaCL alone (row 2) leads to stronger performance upon the baseline (row 1), verifying our hypothesis in Sec. 3.3. Adding false negative filtering (row 4) further boosts the performance, particularly in oIoU. Augmenting motion-centric expressions alongside original captions (row 5) further boosts the performance, guiding the model to better align with actions. Applying all ideas at the same time (row 6) completes our proposed method, achieving the strongest performance.

Ambiguity Filtering Strategy. Our data augmentation strategy exposes the model to more action-centric expressions by extracting verb phrases from original captions. However, ambiguity arises when multiple objects in the same category perform similar actions within a single image. To mitigate this, we apply a pre-filtering strategy that retains only training samples where the target's category appears once. Tab. 7 validates this choice: the application of stricter filtering thresholds consistently improves performance, and the strictest setting (Exclude > 1) achieves the best results in both validation and test sets. This confirms that reducing intra-class ambiguity produces a cleaner training signal for learning motion semantics.

4.4 Effects of Hyperparameters

Margin (m). Tab. 6 reports performance under varying margin values with fixed $\alpha = 0.1$ and $\nu = 0.5$. A margin that is too small fails to enforce sufficient separation between positive and negative samples, while an excessively large margin causes representation collapse and degrades performance. We find that a moderate margin of $m = 12$ achieves the best results, providing an effective balance for contrastive alignment.

Filtering Option	mIoU		oIoU	
	val	test	val	test
No Filtering	58.96	57.98	56.71	57.66
Exclude > 3	59.26	59.96	57.21	57.95
Exclude > 2	60.54	60.38	57.83	58.39
Exclude > 1 (Ours)	**60.87**	**60.47**	**58.05**	**58.95**

Table 7: **Ablation on ambiguity filtering.** The strictest threshold achieves the best results on G-Ref.

Filtering Threshold (ν). As discussed in Sec. 3.2, we filter out potential false negatives during contrastive training. Tab. 6 shows that a moderate threshold around 0.5 results in consistent improvement, likely by eliminating noisy negatives that exhibit high similarity with the anchor. Higher thresholds are less effective as fewer false negatives are detected, diminishing the benefit of filtering. This trend is further supported by Tab. 5, where the optimal threshold leads to clear gains when combined with verb-centric supervision (row 4 vs. row 5).

Metric Loss Weight (α). We vary the metric loss weight α, which determines the relative contribution of $\mathcal{L}_{\text{MRaCL}}$ compared to the segmentation loss $\mathcal{L}_{\text{seg}}$, with fixed $m = 12$ and $\tau = 0.07$. As seen in Tab. 6, $\alpha = 0.1$ achieves the best performance, where larger weights overemphasize the metric loss, slightly dropping the performance.

5 Conclusion

In this work, we address the limitation of existing Referring Image Segmentation (RIS) models in understanding motion-centric expressions. Specifically, we propose to leverage augmented action-centric queries to supplement training examples, and introduce **M**ultimodal **Ra**dial **C**ontrastive **L**oss (**MRaCL**), which performs contrastive learning on fused multimodal embeddings. By utilizing angular distance as the similarity metric, MRaCL overcomes gradient saturation and anisotropy issues inherent in cosine-based methods. Our method leads to superior performance across various RIS models and diverse benchmarks, with particularly significant improvements on test examples requiring a deeper understanding of motions. These results highlight the effectiveness of our method in enhancing both linguistic and visual comprehension for RIS, providing a robust framework for tackling motion-oriented scenarios. As a future work, a similar idea could be applied to the Referring Video Object Segmentation (RVOS) task, requiring even finer-grained understanding of actions.

Acknowledgments

This work was also supported by the SOFT Foundry Institute at SNU, Samsung Electronics, Youlchon Foundation, National Research Foundation of Korea (NRF) grants (RS-2021-NR05515, RS-2024-00336576, RS-2023-0022663, RS-2025-25399604, RS-2024-00333484), and the Institute for Information & Communication Technology Planning & Evaluation (IITP) grants (RS-2022-II220264, RS-2024-00353131) funded by the Korean government.

References

Athar, A., Deng, X., and Chen, L.-C. (2024). ViCaS: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. *arXiv:2412.09754*.

Chen, D.-J., Jia, S., Lo, Y.-C., Chen, H.-T., and Liu, T.-L. (2019). See-through-text grouping for referring image segmentation. In *ICCV*.

Chen, J., Wei, F., Zhao, J., Song, S., Wu, B., Peng, Z., Chan, S.-H. G., and Zhang, H. (2024). Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv:2406.16866*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *ICML*.

Chen, T., Sun, Y., Shi, Y., and Hong, L. (2017). On sampling strategies for neural network-based collaborative filtering. In *KDD*.

Cho, Y., Yu, H., and Kang, S.-J. (2023). Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. *IEEE Transactions on Multimedia*.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *CVPR*.

Chuang, Y.-S., Dangovski, R., Luo, H., Zhang, Y., Chang, S., Soljačić, M., Li, S.-W., Yih, W.-t., Kim, Y., and Glass, J. (2022). DiffCSE: Difference-based contrastive learning for sentence embeddings. *arXiv:2204.10298*.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Ding, H., Liu, C., Wang, S., and Jiang, X. (2021). Vision-language transformer and query generation for referring segmentation. In *ICCV*.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, ELMo, and GPT-2 embeddings. *arXiv:1909.00512*.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. (2019). Representation degeneration problem in training natural language generation models. *arXiv:1907.12009*.

Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. *arXiv:2104.08821*.

Gavrilyuk, K., Ghodrati, A., Li, Z., and Snoek, C. G. (2018). Actor and action video segmentation from a sentence. In *CVPR*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv:2407.21783*.

Ha, S., Kim, C., Kim, D., Lee, J., Lee, S., and Lee, J. (2024). Finding NeMo: Negative-mined mosaic augmentation for referring image segmentation. In *ECCV*.

Hu, R., Rohrbach, M., and Darrell, T. (2016). Segmentation from natural language expressions. In *ECCV*.

Hu, Y., Wang, Q., Shao, W., Xie, E., Li, Z., Han, J., and Luo, P. (2023). Beyond one-to-one: Rethinking the referring image segmentation. In *ICCV*.

Hu, Z., Feng, G., Sun, J., Zhang, L., and Lu, H. (2020). Bi-directional relationship inferring network for referring image segmentation. In *CVPR*.

Huang, J., Xu, Z., Liu, T., Liu, Y., Han, H., Yuan, K., and Li, X. (2025). Densely connected parameter-efficient tuning for referring image segmentation. In *AAAI*.

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

Huynh, T., Kornblith, S., Walter, M. R., Maire, M., and Khademi, M. (2022). Boosting contrastive self-supervised learning with false negative cancellation. In *WACV*.

Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., and Tan, T. (2021). Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. (2021). Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. (2014). ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*.

Kim, N., Kim, D., Lan, C., Zeng, W., and Kwak, S. (2022). ReSTR: Convolution-free referring image segmentation using transformers. In *CVPR*.

Lee, H., Lee, J., Ng, J. Y.-H., and Natsev, P. (2020). Large scale video representation learning via relational graph clustering. In *CVPR*, pages 6807–6816.

Lee, J., Abu-El-Haija, S., Varadarajan, B., and Natsev, A. (2018). Collaborative deep metric learning for video understanding. In *KDD*, pages 481–490.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. (2020). On the sentence embeddings from pre-trained language models. *arXiv:2011.05864*.

Li, R., Li, K., Kuo, Y.-C., Shu, M., Qi, X., Shen, X., and Jia, J. (2018). Referring image segmentation via recurrent refinement networks. In *CVPR*.

Li, X. and Li, J. (2024). AoE: Angle-optimized embeddings for semantic textual similarity. In *ACL*.

Liu, C., Ding, H., and Jiang, X. (2023a). GRES: Generalized referring expression segmentation. In *CVPR*.

Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., and Yuille, A. (2017a). Recurrent multimodal interaction for referring image segmentation. In *ICCV*.

Liu, J., Liu, J., Wang, Q., Wang, J., Wu, W., Xian, Y., Zhao, D., Chen, K., and Yan, R. (2023b). RankCSE: Unsupervised sentence representations learning via learning to rank. *arXiv:2305.16726*.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017b). SphereFace: Deep hypersphere embedding for face recognition. In *CVPR*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

Ma, X., Santos, C. N. d., and Arnold, A. O. (2021). Contrastive fine-tuning improves robustness for neural rankers. *arXiv:2105.12932*.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *CVPR*.

Margffoy-Tuay, E., Pérez, J. C., Botero, E., and Arbeláez, P. (2018). Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*.

Nguyen-Truong, H., Nguyen, E.-R., Vu, T.-A., Tran, M.-T., Hua, B.-S., and Yeung, S.-K. (2025). Vision-aware text features in referring image segmentation: From object understanding to context understanding. In *WACV*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

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

Shi, H., Li, H., Meng, F., and Wu, Q. (2018). Keyword-aware network for referring expression image segmentation. In *ECCV*.

Sohn, K. (2016). Improved deep metric learning with multi-class N-pair loss objective. In *NIPS*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Tang, J., Zheng, G., Shi, C., and Yang, S. (2023). Contrastive grouping with transformer for referring image segmentation. In *CVPR*.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018). CosFace: Large margin cosine loss for deep face recognition. In *CVPR*.

Wang, L., Huang, J., Huang, K., Hu, Z., Wang, G., and Gu, Q. (2019). Improving neural language generation with spectrum control. In *ICLR*.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., and Liu, T. (2022). CRIS: Clip-driven referring image segmentation. In *CVPR*.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *ECCV*.

Yang, J., Zhang, H., Li, F., Zou, X., Li, C., and Gao, J. (2023). Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4v. *arXiv:2310.11441*.

Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., and Torr, P. H. (2022). LAVT: Language-aware vision transformer for referring image segmentation. In *CVPR*.

Ye, L., Rochan, M., Liu, Z., and Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. In *CVPR*.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. (2016). Modeling context in referring expressions. In *ECCV*.

Yu, S., Hong, J., Lee, J., and Son, J. (2025). Latent expression generation for referring image segmentation and grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21374–21383.

Yue, P., Lin, J., Zhang, S., Hu, J., Lu, Y., Niu, H., Ding, H., Zhang, Y., Jiang, G., Cao, L., et al. (2024). Adaptive selection based referring image segmentation. In *ACM MM*.

Zhang, Y., Zhang, R., Mensah, S., Liu, X., and Mao, Y. (2022a). Unsupervised sentence representation via contrastive learning with mixing negatives. In *AAAI*.

Zhang, Y., Zhu, H., Wang, Y., Xu, N., Li, X., and Zhao, B. (2022b). A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *ACL*.

Zhang, Z., Zhu, Y., Liu, J., Liang, X., and Ke, W. (2022c). CoupAlign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. In *NeurIPS*.

Zhou, K., Zhang, B., Zhao, W. X., and Wen, J.-R. (2022). Debiased contrastive learning of unsupervised sentence representations. *arXiv:2205.00656*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**No**]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**] We will release the code if our paper is accepted.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Not Applicable**]
   (b) Complete proofs of all theoretical results. [**Not Applicable**]
   (c) Clear explanations of any assumptions. [**Not Applicable**]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**] Hyperparmeters for the main results are explained in Sec. 4, data preperation processes are explained in App. A and App. C. Code and data will be released on public repository if the paper is accepted.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**]. Details are explained in Sec. 4.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**]. Our method was implemented upon the baselines in Sec. 2 and Sec. 4.1.
   (b) The license information of the assets, if applicable. [**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**]. We plan to provide open access to our MBench evaluation dataset via a public repository if paper is accepted.
   (d) Information about consent from data providers/curators. [**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**No**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

Towards Motion-aware Referring Image Segmentation: Supplementary Materials

A M-Ref Motion & Static Split Examples

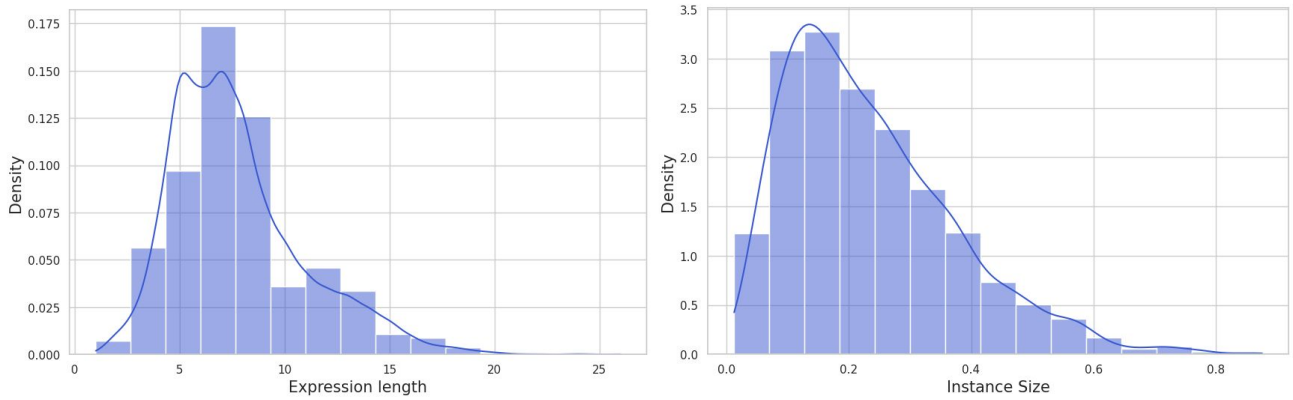
We construct a manually annotated split of the G-Ref UMD test set by categorizing each referring expression as either static or motion-centric. Static queries primarily describe appearance-related or position-based attributes—such as clothing color, physical traits, or spatial relations (*e.g.*, "a man in a baseball uniform", "a person with the thick beard, glasses and a hat")—which leverage visually stable cues and require minimal reasoning about motion. In contrast, motion-centric queries emphasize the dynamic behavior or transient actions of the target object, including expressions like "swinging a bat", "bending over and touching the foot of a horse", or "holding a tennis racket while balancing a ball". While both types of expressions may refer to the same visual target, motion-oriented queries provide more explicit cues about the object's ongoing action, demanding finer understanding and contextual interpretation from the model. This curated split reveals the inherent linguistic and semantic gap between static and dynamic references, serving as a diagnostic benchmark to evaluate how well RIS models can handle both appearance-based and motion-aware expressions. Fig. I illustrates a few examples of image captions in both motion and static splits.

Figure I: Examples of motion and static descriptions in M-Ref.

B MBench Dataset Details

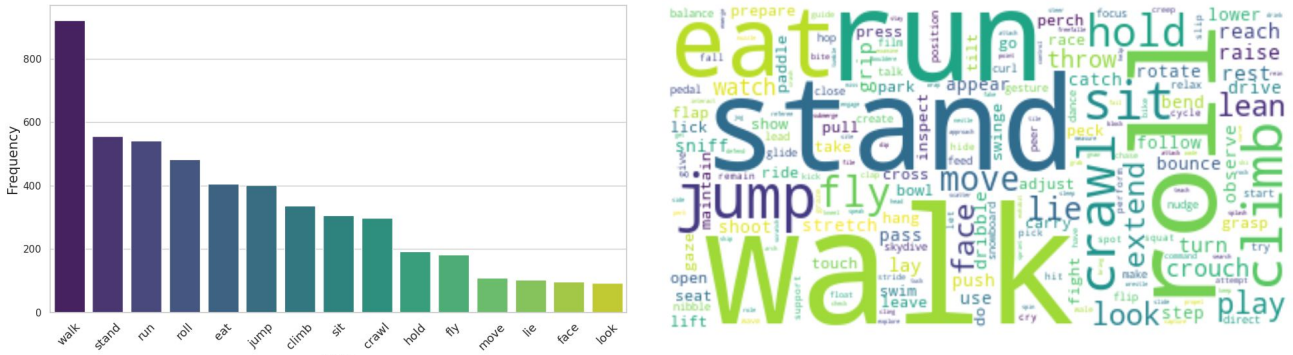
Our MBench consists of image-text pairs where the target instance is differentiated from other objects of the same category using an action-centric expression. It contains 8,200 instance annotations with referring expressions of varying lengths. Each expression contains 7.72 words on average and up to 26 words. Following (Chen et al., 2024), we compute the normalized instance size relative to the image size as

$$\sqrt{\frac{\text{instance area}}{\text{image area}}}, \tag{i}$$

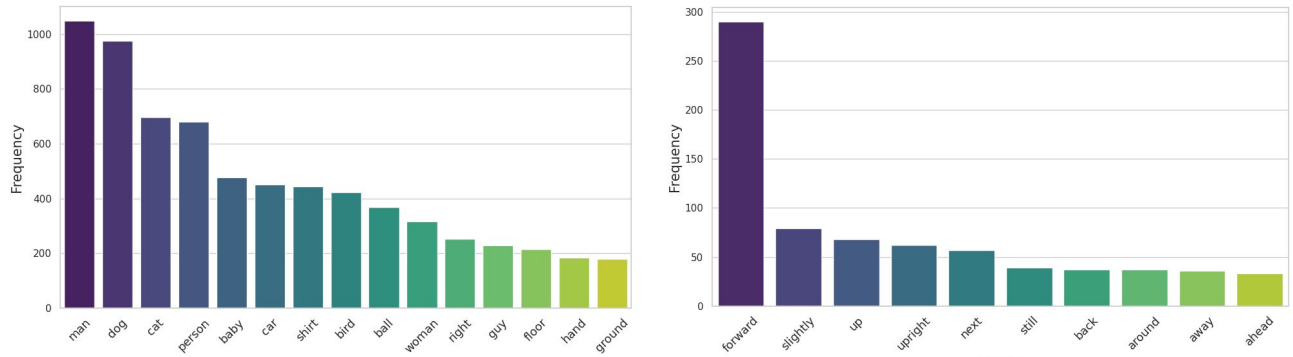


(a) Expression length (word count) (b) Instance Size

Figure II: Analysis of expression length and instance size



(a) Verb count (b) Wordcloud of action expressions



(c) Noun count (d) Adverb count

Figure III: Most frequent verbs, nouns and adverbs

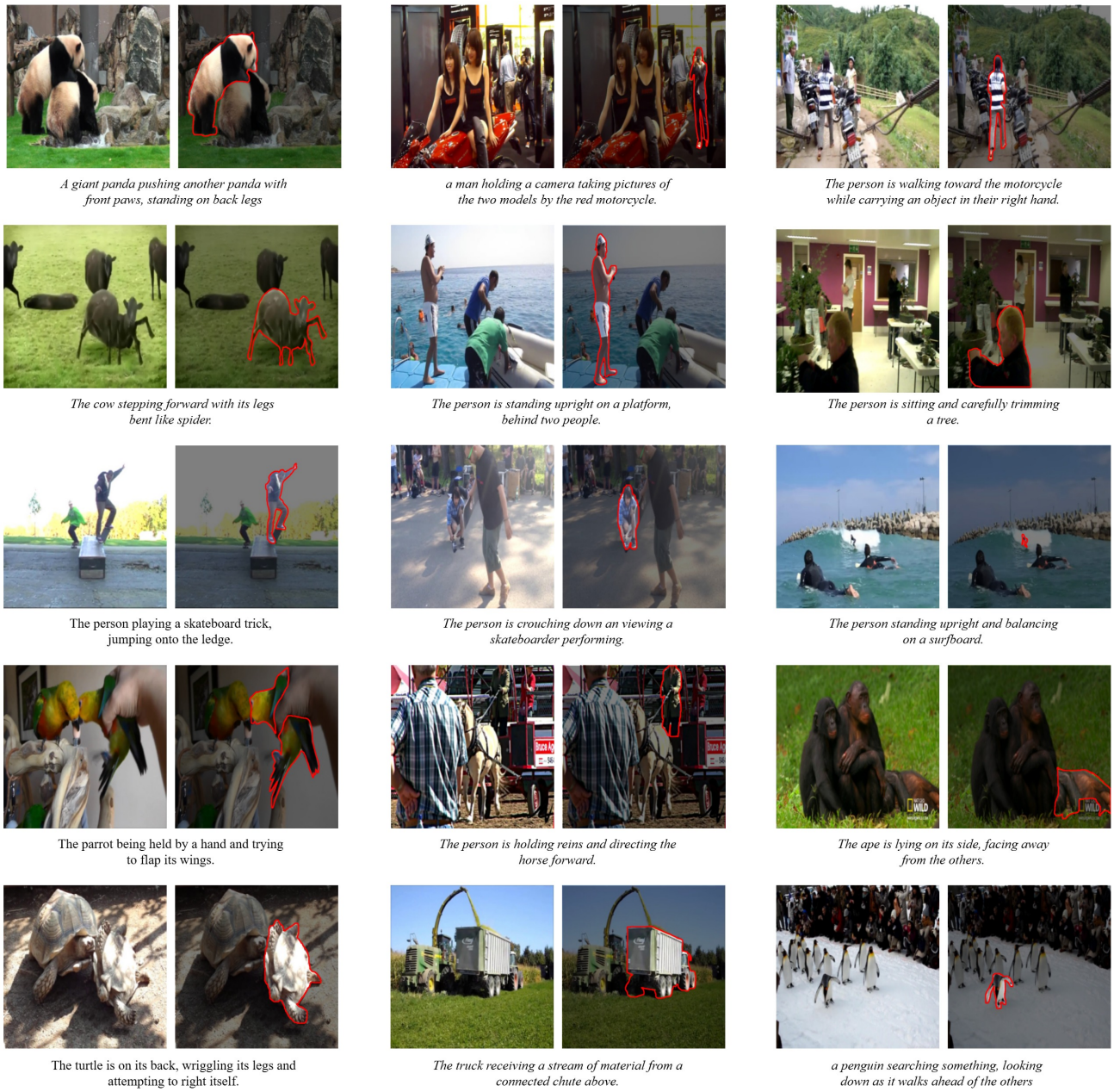


Figure IV: Visualization of MBench image and its highlighted mask

and its distribution, illustrated in Fig. II, peaks at 0.137. The distribution exhibits a long tail, with a maximum value of 0.875, indicating that our dataset targets instances of various sizes. Analyzing the referring expressions using the `spaCy` library, we identify 1,268 unique lemmatized words. The 15 most frequent verbs are *walk, stand, run, roll, eat, jump, climb, sit, crawl, hold, fly, move, lie, face, look*, suggesting that our dataset focuses on casual, everyday activities. We plot the frequencies of 15 most frequent verbs and nouns, and the 10 most frequent adverbs in Fig. III.

In Fig. IV, we illustrate a few examples of the original images and their highlighted masks of MBench. The top-left image shows two giant pandas doing different actions and the model is expected to focus on the visual aspect of 'pushing' and 'standing' and understand the relationship between pandas (*e.g.*, 'pushing another panda') to correctly find the left panda. Action-centric expressions also include passive verbs such as 'being held'. Using these referring expressions, we aim to segment not only animate category objects (*e.g.*, 'person', 'cow', 'ape') but also inanimate objects (*e.g.*, 'truck'). Targeted instances are of various sizes, with smaller objects such as the third image in the right column ("The person standing upright and balancing on a surfboard") of Fig. IV being harder to find. Instances may be partially occluded like the fourth image in the center column ("The person is holding reins and directing the horse forward") and other objects belonging to the same category may be numerous such as the bottom right image ("a penguin searching something, looking down as it walks ahead of the others") making the task especially difficult.

C Details on MBench Construction

Although recently developed RIS datasets contain complicated examples, most of them are still distinguished by appearance cues. In order to evaluate the performance of RIS models with respect to their understanding of action-centric cues in referring expressions, we develop a custom evaluation benchmark called **MBench**, shortened from 'motion expression centric benchmark'. MBench provides a targeted benchmark to assess the ability of a model trained on conventional RIS datasets on action-oriented expressions. By focusing on descriptions where visual disambiguation stems from motion or behavior rather than appearance, MBench serves as a controlled benchmark for assessing the transferability of verb-centric grounding capabilities.

In order to leverage rich action-related cues, we build upon three text-grounded video segmentation datasets; namely, A2D-Sentences (Gavrilyuk et al., 2018), Ref-Youtube-VOS (Seo et al., 2020), and ViCaS (Athar et al., 2024).

Frame Sampling and Category-based Filtering. We first divide each video into four equal-length segments, and sample one frame from each of them to ensure diversity. This step is skipped for A2D, since each annotation already corresponds to a preselected frame.

Then, we annotate a referring expression for each object instance in the selected frames, describing its visual context and behavior within the scene. Considering our goal to evaluate action-centric disambiguation, it is important that multiple co-occurring instances in the scene are distinguishable by their behavior. Thus, we keep frames where at least two or more instances belonging to the same category (*e.g.*, woman, horse, cat) cannot be trivially differentiated by category name alone.

If a frame contains only one instance of a given category, it is excluded from the benchmark. For this step, we use the category annotations if provided in the base set, where each annotated instance is labeled with a corresponding target category ID. Otherwise, we parse the main noun phrase from the referring expression as a pseudo-category. For example, from a caption like "A man in shorts is exercising with a rubber band", the parser derives candidate labels such as "man in shorts" and "rubber band". These extracted phrases serve as semantic object categories and are used for downstream filtering and evaluation.

We then determine whether each identified category belongs to a movable entity (*e.g.*, human, animal, vehicle) based on a manually curated list. In cases where pseudo-categories are used, we query a lightweight language model (Gemini (Team et al., 2023) 2.0 Flash Lite) with prompts like "Is {category} capable of independent motion or typical actions?" to assess movability. Only objects confirmed as movable are retained.

Caption Generation. For each frame that satisfies the criteria above, we proceed to the captioning stage. First, we overlay the segmentation masks provided by each dataset as unfilled polygons on the image. Following Yang et al. (Yang et al., 2023), we additionally annotate an object ID at the center of each instance to facilitate

clear identification, as illustrated in Fig. 4. This setup ensures that all instances within a frame are distinctly marked, allowing unambiguous reference and precise disambiguation during captioning.

Each marked frame is then fed into an LLM (*e.g.*, we use GPT-4o) via a two-stage 'visual chain-of-thought prompting' strategy. In the first stage, we verify if the identified objects of the same category are performing distinct and recognizable actions (Movable Action Filtering). If so, we proceed to the second stage, where we prompt the LLM to generate dense object-level captions focusing solely on the prominent action verb, excluding appearance, spatial relations, or ambiguous speculations. We utilize multiple prompt templates to encourage diversity in expression.

Text Refinement. We observe that the initially generated captions often contain noisy patterns, such as explicit instance IDs (*e.g.*, "the giant panda marked '2' "), redundant action descriptions, or unnecessary appearance expressions. To improve clarity and consistency of the captions, we rewrite them into concise action descriptions focused on individual objects, using a lightweight language model (Llama-3.1-8B (Grattafiori et al., 2024)). This step ensures action-oriented queries, aligned with our goal of producing this benchmark set.

Lastly, we manually review a subset of the resulting data to construct a high-quality validation set, discarding samples if their generated captions fail to reflect the object's actual behavior or if multiple instances exhibit indistinguishable actions. Annotators are presented with the marked frame and the generated caption, and are asked to verify both the semantic accuracy and visual alignment of the descriptions. Through this pipeline, we obtain 8,200 action-focused evaluation samples (instances) for the RIS task, requiring fine-grained object understanding in dynamic visual scenes. We will publicly release this benchmark set upon acceptance.

D Additional result on more advanced baseline

We additionally compare with a recently published state-of-the-art RIS model Latent-VG (Yu et al., 2025). As shown in the table below, we observe consistent performance improvements across all three datasets, aligning with the trends reported in Tab. 2 (oIoU scores). For a fair comparison, we reproduce all baselines under the same setting as noted in Sec. 4.1.

Table I: **Additional results with LatentVG (Yu et al., 2025) as baseline (oIoU).** Applying MRaCL to a more recent and stronger baseline yields consistent improvements across all splits of all three datasets.

Method	**RefCOCO (UNC)**			**RefCOCO+ (UNC+)**			**G-Ref (UMD)**	
	Val	**TestA**	**TestB**	**Val**	**TestA**	**TestB**	**Val**	**Test**
Latent-VG (reproduced)	76.23	78.39	73.33	69.16	72.83	62.18	69.07	69.83
Latent-VG + MRaCL	**76.94**	**78.87**	**73.95**	**69.49**	**73.51**	**62.56**	**70.02**	**70.66**

E More Qualitative Results

We present additional qualitative results of applying our framework to the state-of-the-art baselines, evaluated on the MBench dataset. As shown in Fig. V, models trained with our method demonstrate consistent improvements in segmenting target entities described by action-centric referring expressions. These qualitative examples further illustrate how our proposed method enables RIS models to capture nuanced verb-centric semantics without requiring full model retraining, in line with the goals outlined in our approach.

For example, in the top row of Fig. V, the expression "*The cat that is not running, but standing at the side*" requires grounding a subtle, non-salient action. Baselines often misidentify the subject, whereas those equipped with our method correctly localize the stationary cat by capturing fine-grained pose semantics. In the second case, the expression "*The seal lying on its belly and raising its flapper to reach*" involves two concurrent actions. Without our framework, baseline models fail to fully segment the active region or confuse nearby objects. In contrast, models trained with ours successfully align the dual motions—body posture and limb movement—demonstrating improved understanding of compositional, action-centric expressions.

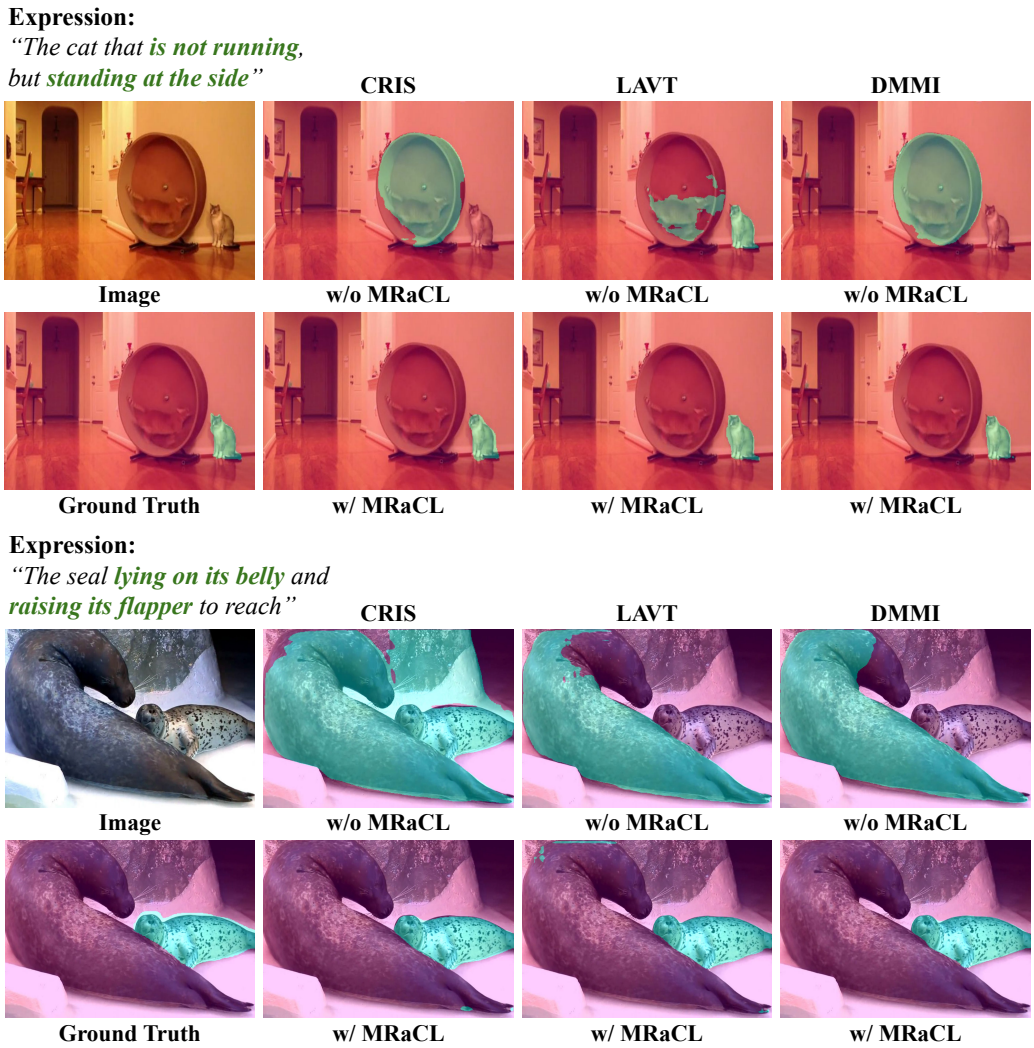


Figure V: **Additional visualizations of predictions from the MBench test set.** For each example, the left-most column presents the input expression, followed by the input image, and the ground truth mask overlaid on the image for reference. Subsequent columns across each row exhibit the comparative visualizations of model predictions before and after applying our MRaCL based framework.