Semantic-conditioned Dual Adaptation for Cross-domain Query-based Visual Segmentation

Ye Wang[*], **Tao Jin**[*], **Wang Lin**[*],
Xize Cheng, **Linjun Li**, **Zhou Zhao**[†]
Zhejiang University
{yew,jint_zju,linwanglw}@zju.edu.cn
{chengxize,lilinjun21,zhaozhou}@zju.edu.cn

Abstract

Visual segmentation from language queries has attracted significant research interest. Despite the effectiveness, existing works require expensive labeling and suffer severe degradation when deployed to an unseen domain. In this paper, we investigate a novel task Cross-domain Query-based Visual Segmentation (CQVS), aiming to adapt the segmentation model from a labeled domain to a new unlabeled domain. The challenges of CQVS stem from three domain discrepancies: (1) multi-modal content shift, (2) uni-modal feature gap and (3) cross-modal relation bias. Existing domain adaptation methods fail to address them comprehensively and precisely (e.g. at pixel level), thus being suboptimal for CQVS. To overcome this limitation, we propose Semantic-conditioned Dual Adaptation (SDA), a novel framework to achieve precise feature- and relation-invariant across domains via a universal semantic structure. The SDA consists of two key components: Content-aware Semantic Modeling (CSM) and Dual Adaptive Branches (DAB). First, CSM introduces a common semantic space across domains to provide uniform guidance. Then, DAB seamlessly leverages this semantic information to develop a contrastive feature branch for category-wise pixel alignment, and design a reciprocal relation branch for relation enhancement via two complementary masks. Extensive experiments on three video benchmarks and three image benchmarks evidence the superiority of our approach over the state-of-the-arts.[1]

1 Introduction

Vision-language understanding (Yin et al., 2022, 2021; Jin and Zhao, 2021b; Jin et al., 2022; Cheng et al., 2023) is a fundamental problem in deep learning. Recently, in this field, query-based visual segmentation (Wang et al., 2020; Botach et al., 2022) has received considerable critical attention, which aims at localizing the visual region at the pixel-level that semantically corresponds to the language query. Though existing works have made tremendous progress, the manual collection of pixel-wise annotations is expensive and tedious in practice, raising a significant requirement for the application in unseen circumstances. However, varying construction conditions on different domains inevitably degrades the adaptation performance, which is formally known as the *domain shift* problem.

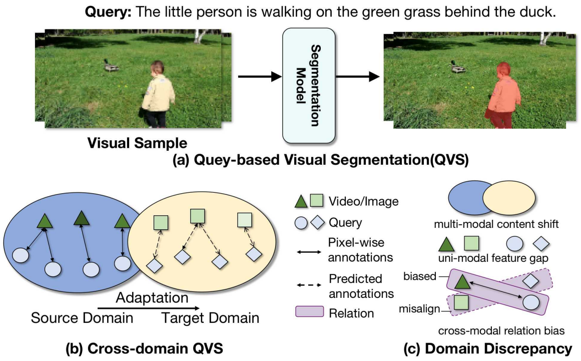


Figure 1: (a) Example of the query-based visual segmentation. (b) Illustration of the cross-domain query-based visual segmentation. (c) Three domain discrepancies.

In this paper, to breakthrough the constraint, we propose a novel task Cross-domain Query-based Visual Segmentation (CQVS). As shown in Figure 1, given a source domain with pixel-wise annotations and an unlabeled target domain, CQVS aims to adapt the segmentation model and recognize the query-aligned pixels on the target domain.

To achieve effective adaptation for this cross-modal grounding task, we have to deal with three domain discrepancies as shown in Figure 1 (c): (1) Multi-modal content shift. The free-form query and aligned visual region describe open and diverse contents, leading to arbitrary semantic shift between domains, e.g. one domain mainly describes humans while the other focus more on animals. (2)

[*] Equal contribution.

[†] Corresponding author

[1] https://github.com/yewzz/SDA.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9797–9815
July 9-14, 2023 ©2023 Association for Computational Linguistics

Uni-modal feature gap. Even describing the same content, each modality may have huge feature gap between domains caused by varying conditions, e.g. visual light and linguistic syntax. (3) Cross-modal relation bias. The relation between modalities are easily learned to be biased by domain-specific factors, especially when only source annotations is available. As illustrated, it cannot be guaranteed to be aligned across domains by separately aligning each modality and requires dedicated solutions.

To mitigate domain discrepancies, domain adaptation (DA) methods align the distributions across domains (Baktashmotlagh et al., 2014) or learn domain-invariant representations (Ganin et al., 2016). While promising, they are confined for CQVS by two critical limitations. (1) Traditional DA methods study the uni-modal tasks, e.g. image segmentation (Vu et al., 2019) and text classification (Glorot et al., 2011), which are insufficient without consideration of multi-modal content and cross-modal relation. (2) Though recent works (Jing et al., 2020; Liu et al., 2021b) investigate the multi-modal tasks, they (i) are limited to image-level retrieval that is imprecise compared with pixel-level grounding, and (ii) only consider partial discrepancies (feature and relation) regardless of internal correlation between domain contents. Motivated by the fact that humans can leverage abstract concepts to guide thinking about concrete details (Thomas and Thorne, 2009), we aim to model the high-level semantic structure in multimodal contents to harmonize the low-level pixel adaptation for feature and relation discrepancies.

Grounded on the above discussions, we propose Semantic-conditioned Dual Adaptation (SDA), a novel framework to achieve precise feature- and relation-level adaptation via a universal semantic structure. Our proposed SDA consists of two key components, Content-aware Semantic Modeling (CSM) and Dual Adaptive Branches (DAB). (1) CSM builds a sharable semantic space across domains based on the multi-modal content. First, to discover the consistent contents from visual and textual modalities, we extract informative words with visual-guided attention. Then we establish the semantic structure upon the contents, where we apply unsupervised clustering on the source domain and measure the cross-domain semantic similarity to identify common parts on the target domain. Through this module, we implicitly encode the language gap and introduce semantic category information as common guidance to regularize adaptation. (2) Next, DAB seamlessly integrates dual branches to separately address feature and relation discrepancies. On the one hand, a contrastive feature branch leverages the semantic information to learn category-wise alignment for foreground pixels. To preserve pixel-level discrimination during aligning, we adopt contrastive learning (He et al., 2020; Jin and Zhao, 2021a) to highlight relevant foreground pixels between domains and suppress diverse background pixels. On the other hand, a reciprocal relation branch mitigates the cross-modal relation bias via two reciprocal masks. Concretely, the domain-based masks induced by source knowledge can provide precise but biased results and the semantic-based masks induced by semantic information can provide comprehensive but inaccurate results. With the complementary signals, we indirectly enhance relation via segmentation training on both domains and also directly maximize the mutual information between vision and language.

In summary, our main contributions are listed as follows: (1) We propose a new task CQVS to explore domain adaptation for query-based visual segmentation. (2) We introduce a novel framework SDA, which develops a content-aware semantic modeling module to model the multimodal contents and designs a dual adaptive branches module to mitigate the feature and relation discrepancies. (3) Extensive experiments on both query-based image and video segmentation datasets evidence the effectiveness and superiority of our approach.

2 Related Work

2.1 Query-based Visual Segmentation

Query-based visual segmentation aims at recognizing the relevant pixels in the images or videos based on the language query. For image segmentation, several works explore cross-modal fusion methods (Hu et al., 2016; Liu et al., 2017; Margffoy-Tuay et al., 2018), extract multi-modal context with attention (Ye et al., 2019; Huang et al., 2020; Jain and Gandhi, 2021), develop cycle-consistency learning (Chen et al., 2019b) and study the adversarial training (Qiu et al., 2019). Recently, more works investigate video segmentation. Several works focus on dynamic convolution-based methods (Gavrilyuk et al., 2018; Wang et al., 2020; Hui et al., 2021), explore cross-modal attention (Wang et al., 2019; Liu et al., 2021a) and study visual-textual capsule routing algorithms (McIntosh et al., 2020).

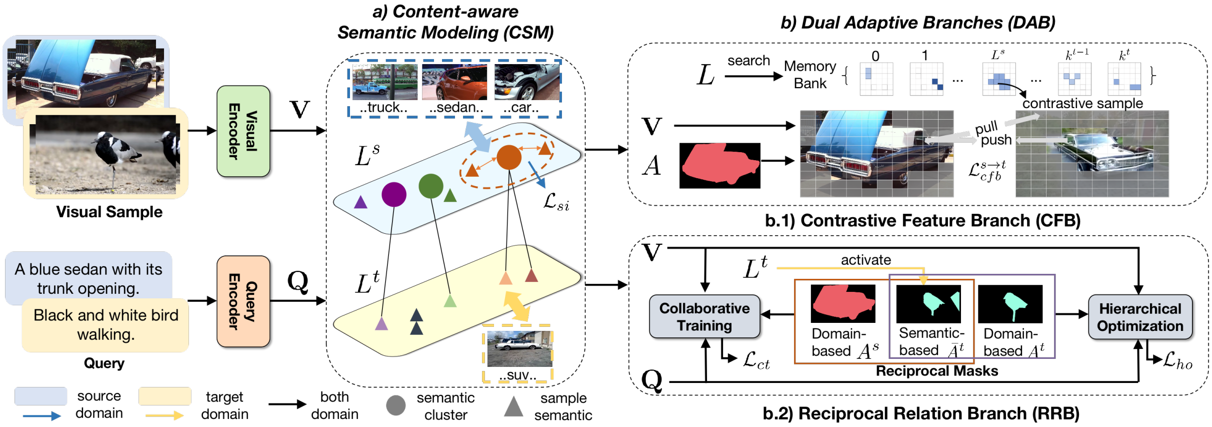


Figure 2: The illustration of the proposed Semantic-conditioned Dual Adaptation framework.

Recent work (Botach et al., 2022) builds the transformer model for this task. However, existing research heavily relies on expensive annotations and could barely generalize to unseen circumstances, hindering the feasibility in practice. Hence, we further study the domain adaptation for this task.

2.2 Domain Adaptation

Traditional unsupervised domain adaptation (UDA) methods have been explored to close the domain gap, e.g., maximum mean discrepancy (Long et al., 2017, 2018) and adversarial learning (Ganin and Lempitsky, 2015; Chen et al., 2019a). In the visual research, more advanced approaches are investigated, e.g., image classification (Ganin and Lempitsky, 2015; Pan et al., 2020), semantic segmentation (Vu et al., 2019; Zhang et al., 2021) and object detection (Saito et al., 2019; Li et al., 2022). However, these methods merely consider the domain discrepancy of the single modality.

Recently, few works study domain adaptation for multi-modal tasks, e.g., image captioning (Chen et al., 2017), text-based person search (Jing et al., 2020) and video-text retrieval (Chen et al., 2021; Liu et al., 2021b). Despite the effectiveness, they fail to comprehensively and precisely address various domain discrepancies in this cross-modal segmentation task. Thus, we propose a novel framework to achieve effective adaptation for CQVS.

3 Method

3.1 Preliminary

Problem Formulation. In this task, we are given a labeled source domain $\mathcal{D}^s = \{V_i^s, Q_i^s, A_i^s\}_{i=1}^{N_s}$ and an unlabeled target domain $\mathcal{D}^t = \{V_i^t, Q_i^t\}_{i=1}^{N_t}$ containing N_s and N_t samples respectively, where V, Q, A is the visual sample (*i.e.* image/video), textual query and pixel-wise annotations. The goal is to construct a model with existing data to segment the query-relevant pixels on the target domain $\mathcal{D}^t$.

Base Network. To illustrate our SDA paradigm clearly, we first formulate the base segmentation network as three common modules: (1) **Encoder**: Given the raw visual input V and query embedding $\mathbf{W}$, it encodes visual features as $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times C}$ and query features as $\mathbf{Q} \in \mathbb{R}^{N\times C}$, where T,H,W,C are the frame number, height, width and hidden size of the visual features respectively, N is the word number. Note $T = 1$ for the image and T will be omitted for ease of presentation. Based on annotations A^s, the visual features $\mathbf{V}^s$ can be correspondingly divided into the foreground $\mathbf{V}^{s,+}$ and the background $\mathbf{V}^{s,-}$; (2) **Interaction**: It develops the cross-modal interaction with attention weight $\alpha \in \mathbb{R}^{H\times W\times N}$ and outputs enhanced pixel-level features $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$; (3) **Decoder**: It is applied on $\mathbf{F}$ to generate final response map $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^{H\times W}$. To train the model, we utilize a binary cross-entropy loss as $\mathcal{L}_{seg} = \mathcal{L}_{bce}(\mathbf{S}, A)$.

Overall Framework. The overall SDA is shown in Figure 2. SDA includes a CSM module (Section 3.2) that encodes the content shift and a DAB module (Section 3.3) that closes the feature- and relation-level gap, achieving effective adaptation.

3.2 Content-aware Semantic Modeling

To explore the semantic structure of open and diverse vision-language content, we propose a novel method to extract informative contents via cross-modal attention, and construct a universal semantic space by leveraging a clustering algorithm and measuring cross-domain semantic similarity.

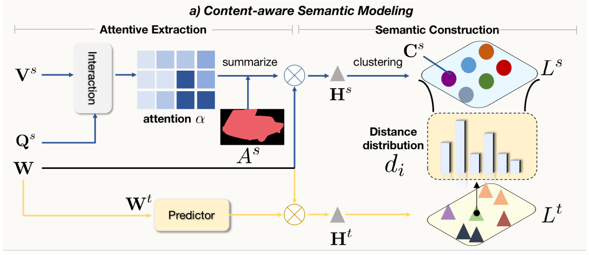


Figure 3: The illustration of the CSM module.

Attentive Extraction. The multi-modal contents exist in the informative and common parts of visual and textual modalities. As visual features easily vary under different conditions and textual features are comparatively stable but scattered over the sequence, we employ vision-language attention to highlight the correlated words in the query that are attended by the visual foreground. In the source domain, with attention $\alpha^s \in \mathbb{R}^{H\times W\times N}$ in the interaction module and the annotations $A^s \in \{0,1\}^{H\times W}$, we calculate visual-guided attention $\{\bar{\alpha}_n^s\}_{n=1}^N$ over words by averaging the attention scores of all foreground pixels as $\bar{\alpha}_n^s = \frac{1}{H\times W}\sum_{i=1}^{H\times W} \alpha_{i,n}^s A_i^s$. Then we normalize $\{\bar{\alpha}_n^s\}_{n=1}^N$ with temperature τ for a sharper distribution and combine it with query embeddings $\mathbf{W} = \{\mathrm{w}_i\}_{i=1}^N$ to obtain content features $\mathbf{H}^s \in \mathbb{R}^C$, given by:

$$\tilde{\alpha}_n = \frac{\exp(\bar{\alpha}_n/\tau)}{\sum_{n=1}^N \exp(\bar{\alpha}_n/\tau)},\ \mathbf{H}^s = \frac{1}{N}\sum_{n=1}^N \tilde{\alpha}_n^s \mathrm{w}_n^s \quad (1)$$

Here we adopt the embedding $\mathbf{W}$ rather than $\mathbf{Q}$ for its generalizable ability (Bravo et al., 2022).

In the target domain without annotations, we leverage knowledge distillation (Hinton et al., 2015) to train a predictor with attention weights $\bar{\alpha}^s$ as guidance, enabling it to directly learn the importance from the query. Thus, we can adaptively predict attention weights $\bar{\alpha}^t$ and calculate content features $\mathbf{H}^t$. Details are listed in Appendix B.1.

Semantic Construction. On the basis of multi-modal contents, we aim to abstract and summarize the high-level semantics that represents the key concepts. Concretely, we apply Agglomerative Clustering (Zhang et al., 2012) upon source contents $\mathbf{H}^s$ and obtain k^s different semantic categories with their prototypes (i.e. center of cluster) $\{\mathbf{C}_i^s\}_{i=1}^{k^s}$. To explore the target semantic structure, we then measure the cross-domain semantic similarity by calculating the sample-to-prototype distance. That is, we compute the similarity distribution $d_i = \{d_{i,j}\}_{j=1}^{k^s}$ from the semantic $\mathbf{H}_i^t$ of the i-th target sample to all source prototypes $\{\mathbf{C}_i^s\}_{i=1}^{k^s}$, given by:

$$d_{i,j} = \frac{\exp(\mathbf{H}_i^{t\top}\mathbf{C}_j^s)}{\sum_{n=1}^{k^s}\exp(\mathbf{H}_i^{t\top}\mathbf{C}_n^s)} \quad (2)$$

With the cross-domain semantic similarity d_i, we can define the boundary ρ between "common" and "unknown" points based on its entropy $\mathcal{H}(d_i)$, so as to align the common parts of the target domain to the source domain and reject unknown parts. Each reliable target sample with entropy below the boundary ρ will be assigned to the nearest source cluster while other unreliable ones will be clustered into extra classes, resulting in $k^t (\leq k^s)$ common and k^u novel classes. With semantic category labels L assigned to data pairs, we establish a semantic structure across domains, where each source category label $L_i^s \in \{1, 2, .., k^s\}$ and each target category label $L_j^t \in \{1, 2, .., k^t, .., k^t + k^u\}$.

Semantic Initialization. To further initialize discriminative pixel features for stable adaptation, we retrain the segmentation model and incorporate the category clue to ensure the foreground features compact within categories and separate between categories. With the category L^s, we can similarly obtain visual prototypes $\{\mathbf{U}_i^s\}_{i=1}^{k^s}$ for foreground pixels $\mathbf{V}^{s,+}$. Following the work (Zhang et al., 2021), we calculate the similarity between $\{\mathbf{U}_i^s\}_{i=1}^{k^s}$ and $\mathbf{V}^{s,+}$ as $\mathrm{P}^s = [\mathrm{P}_1^s, \mathrm{P}_2^s, .., \mathrm{P}_{k^s}^s]$, then compute the prototypical contrastive loss $\mathcal{L}_{si}$ by:

$$\mathrm{P}_i^s = \frac{\exp(\mathbf{V}^{s,+}\mathbf{U}_i^s)}{\sum_{n=1}^{k^s}\exp(\mathbf{V}^{s,+}\mathbf{U}_n^s)}, \mathcal{L}_{si} = \mathcal{L}_{ce}(\mathrm{P}^s, L^s) \quad (3)$$

3.3 Dual Adaptive Branches

In DAB module, we develop two branches to separately address feature and relation discrepancies.

3.3.1 Contrastive Feature Branch

In this branch, we adopt contrastive learning (He et al., 2020) to achieve category-wise foreground alignment for the visual feature gap.

For visual samples from two domains with the same foreground category, we aim to contrastively strengthen their foreground agreement. While source foreground can be obtained via annotations A^s, we apply two thresholds γ_{min} and γ_{max} on the predicted response map $\mathbf{S}^t$ and filter the highly reliable pixels to obtain pseudo masks A^t on the target domain, where the position with score over γ_{max} and below γ_{min} are set to 1 and 0 respectively while the rest are ignored. Then we can similarly obtain foreground and background features $\mathbf{V}^{t,+}$ and $\mathbf{V}^{t,-}$. We calculate the pixel-level similarity from the source foreground $\mathbf{V}^{s,+}$ to the target reliable pixels $\mathbf{V}^{t,*} = \mathbf{V}^{t,+} \cup \mathbf{V}^{t,-}$, and make the similarity of associated foreground pairs higher than any other irrelevant pairs, given by:

$$\mathcal{L}_{cfb}^{s\to t} = -\frac{1}{|G^s|}\sum\nolimits_{i\in G^s} \log\frac{\exp(\mathbf{V}_i^{s,+}\mathbf{V}_j^{t,+})}{\sum_n \exp(\mathbf{V}_i^{s,+}\mathbf{V}_n^{t,*})} \quad (4)$$

where G^s is the set of source foreground pixels.

To enhance the diversity of contrastive samples, we maintain a memory bank $\{\{\mathbf{m}_i^l\}_{i=1}^B\}_{l=1}^{k^t}$ with B feature maps for k^t common categories on each domain, providing abundant samples based on the category and domain of current training data.

The contrastive training is developed bidirectionally to enable the pixel features from each domain can be enhanced by the other. The full loss combines the symmetric terms as $\mathcal{L}_{cfb} = \mathcal{L}_{cfb}^{s\to t} + \mathcal{L}_{cfb}^{t\to s}$.

3.3.2 Reciprocal Relation Branch

In this branch, we learn semantic-based masks as reciprocal signals and enhance relation with two designed modules to alleviate the relation bias.

Reciprocal Masks. Typical methods optimize relation via annotations A^s on the source domain and pseudo masks A^t on the target domain that are refined from the predicted response map $\mathbf{S}^t$ (Section 3.3.1). However, A^t essentially relies on the decoder pre-trained on the source domain. Hence, A^s and A^t are both domain-based masks and suffer bias due to coupling with source knowledge. Instead, we leverage the domain-agnostic semantic category L^t to develop a multi-label visual classification on the target domain and obtain semantic-based masks $\bar{A}^t \in \{0,1\}^{H\times W}$ from the class activation map (Appendix B.2), which can highlight the instances of the specific category. The two types of masks are complementary: (1) A^s provides accurate source annotations for segmentation ability while $\bar{A}^t$ provides independent target masks as external knowledge. (2) A^t focuses on precise but biased pixels on the target domain while $\bar{A}^t$ provides imprecise but comprehensive category instances as reciprocal signals (Figure 9).

Collaborative Training. Given A^s on the source domain and $\bar{A}^t$ on the target domain, direct training with mixed annotations of different granularity is ineffective (Luo and Yang, 2020). Thus, we collaboratively train the model on two domains via a shared encoder and two separate decoders for two annotations respectively, eliminating the effect of inaccurate masks on the main decoder and providing comprehensive information from the shared encoder. With the source output $\mathbf{S}^s$ from the main decoder and the target output $\bar{\mathbf{S}}^t$ from the auxiliary decoder, the objective is given by:

$$\mathcal{L}_{ct} = \mathcal{L}_{bce}(\mathbf{S}^s, A^s) + \mathcal{L}_{bce}(\bar{\mathbf{S}}^t, \bar{A}^t) \quad (5)$$

Hierarchical Optimization. We also enhance relation based on the cross-modal mutual information (MI). On the target domain, with the domain-based and semantic-based masks A^t and $\bar{A}^t$, we denote the features of their intersection part as the selected instance $\mathbf{V}_{sel}^t$, the features of their union part as category instances $\mathbf{V}_{ct}^t$ and the features of the remaining part as background $\mathbf{V}_{bg}^t$. To distinguish the hierarchical confrontment among them (i.e. background-category-instance), we follow the work (Hjelm et al., 2018) to maximize MI by:

$$\begin{aligned}&\mathrm{MI(a,b^+,b^-)} = \mathbb{E}[(\phi(\mathrm{a},\mathrm{b}^+)] - \mathbb{E}[(\phi(\mathrm{a},\mathrm{b}^-))]\\ &\mathcal{L}_{ho}^t = \mathrm{MI}(\mathbf{Q},\mathbf{V}_{sel}^t,\mathbf{V}_{ct}^t) + \mathrm{MI}(\mathbf{Q},\mathbf{V}_{ct}^t,\mathbf{V}_{bg}^t)\end{aligned} \quad (6)$$

where $\phi(\cdot,\cdot)$ is the MI discriminator followed by the softplus function. Similarly, we enhance MI with loss $\mathcal{L}_{ho}^s$ in the source domain by directly distinguishing between the foreground and the background. The final objective $\mathcal{L}_{ho} = \mathcal{L}_{ho}^s + \mathcal{L}_{ho}^t$.

3.4 Training and Inference

Training. We develop a multi-stage training. (1) We pre-train the segmentation model with loss $\mathcal{L}_{seg}$ on the source domain. (2) Then we leverage the pre-trained model in CSM module and re-train it by adding the loss $\mathcal{L}_{si}$. (3) We incorporate the DAB module to continue training with the full loss by:

$$\mathcal{L}_{sda} = \mathcal{L}_{ct} + \lambda_1\mathcal{L}_{cfb} + \lambda_2\mathcal{L}_{ho} \quad (7)$$

Inference. During inference, we use the main decoder to segment pixels with score higher than half of the max value in the response map as foreground.

4 Experiments

4.1 Setup

Datasets. For query-based visual segmentation, we consider three video datasets: **Refer-Youtube-VOS** (Seo et al., 2020), **A2D Sentences** (Gavrilyuk et al., 2018) and **J-HMDB Sentences** (Gavrilyuk et al., 2018). We also evaluate our method on three image datasets: **UNC** (Yu et al., 2016), **UNC+** (Yu

Method	RVOS → A2D			A2D → RVOS			RVOS → J-HMDB			A2D → J-HMDB		
	IoU		mAP	IoU		mAP	IoU		mAP	IoU		mAP
	O	M	0.5:0.95	O	M	0.5:0.95	O	M	0.5:0.95	O	M	0.5:0.95
Base(Ours)	40.21	50.60	21.36	30.24	35.46	13.17	56.42	58.72	26.23	61.48	62.54	35.28
CMDyConv	37.32	46.44	18.36	26.71	32.08	10.12	54.18	56.62	24.07	59.74	60.27	33.62
CMPC-V	40.56	49.83	20.72	29.64	34.92	13.01	57.33	58.61	26.74	61.26	61.40	34.11
MMD	41.19	51.26	22.44	31.06	35.38	14.22	57.13	59.16	27.01	62.04	63.11	36.26
DANN	39.73	49.62	20.14	28.44	34.01	11.96	55.25	57.42	25.64	61.96	62.74	35.71
PLCA	42.03	51.78	22.96	31.87	36.02	14.66	58.23	59.41	28.24	63.12	64.08	36.81
MAN	42.36	51.76	23.84	31.02	35.97	14.21	58.13	60.21	28.56	63.17	64.36	37.42
ACP	43.09	51.92	24.13	31.81	36.24	15.06	59.02	60.79	29.34	62.56	63.28	36.73
SDA	**46.31**	**53.10**	**27.38**	**33.41**	**38.26**	**16.52**	**63.47**	**64.47**	**34.90**	**66.03**	**66.75**	**39.37**

Table 1: Performance comparisons of four transfer tasks on three video datasets. O=Overall. M=Mean.

Method	UNC → UNC+			UNC → G-Ref		
	IoU		mAP	IoU		mAP
	O	M	0.5:0.95	O	M	0.5:0.95
Base(Ours)	34.57	38.31	23.15	37.45	41.09	26.42
CMSA	33.32	36.94	21.61	36.42	40.28	25.06
CMPC-I	35.23	38.98	23.26	38.61	41.74	27.13
MMD	35.44	38.74	23.96	37.96	41.88	27.02
DANN	33.18	37.42	21.17	35.98	39.76	24.88
PLCA	35.62	39.81	24.46	38.64	41.68	28.02
MAN	36.44	40.03	24.82	38.53	42.12	27.66
ACP	36.39	39.44	25.11	38.42	42.43	28.13
SDA	**37.89**	**41.23**	**26.34**	**39.64**	**43.13**	**29.76**

Table 2: Performance comparisons of two transfer tasks on three image datasets.

et al., 2016) and **G-Ref** (Mao et al., 2016). Since the image datasets are mostly collected on MS-COCO (Lin et al., 2014), we conduct more experiments on the challenging video datasets.

Evaluation Metrics. Following prior works, we employ the criterias including IoU (Intersection-over-Union) and mean average precision as metrics. For IoU, we compute the **Overall IoU** and the **Mean IoU**. We compute the mean average precision over different thresholds as **mAP[0.5:0.95]**.

More details of dataset statistics and the implementation details are summarized in Appendix C.

4.2 Performance Comparison

Baselines. We compare SDA with the following methods. (1) For query-based visual segmentation methods that only utilize the labeled source data for training, we select CMDyConv (Wang et al., 2020), CMPC-V (Hui et al., 2021) for videos, and consider CMSA (Ye et al., 2019), CMPC-I (Hui et al., 2021) for images. (2) For DA methods that utilize both the labeled source and unlabeled target data for training, we consider the uni-modal DA methods: MMD (Long et al., 2015), DANN (Ganin and Lempitsky, 2015) for image classification, PLCA (Kang et al., 2020) for semantic segmentation and the cross-modal DA methods: MAN (Jing et al., 2020) for text-based person search, ACP (Liu et al., 2021b) for vision-language retrieval.

Query-based Video Segmentation: The results of four transfer directions on three video datasets are shown in Section 3.4. (1) We observe that our SDA framework consistently outperforms all other methods on all criterias, improving mAP[0.5:0.95] by 6.0, 3.3, 8.7, 4.1 on four transfer tasks respectively. (2) The uni-modal DA method MMD brings little gains and DANN even slightly degrades the performance. We infer the reason is that directly aligning each modality results in negative transfer, as discussed in Section 1. (3) Though cross-modal DA methods achieve a performance boost, they are still inferior to our approach due to the lack of a comprehensive solution to the domain discrepancies. The above observations solidly demonstrate the strong adaptation ability of our SDA framework.

Query-based Image Segmentation: As shown in Section 3.4, our SDA also achieves the best results on two transfer tasks for query-based image segmentation. The fact validates the generalizable ability of our approach on different visual modalities (image/video) and further evidences its effectiveness.

4.3 Ablation Study

To investigate the validity of the derived modules, we conduct ablation studies on two adaptation tasks RVOS → A2D and UNC → UNC+.

Main Ablation Study. As shown in Section 4.3, we verify the contribution of each module in our SDA. The **CSM** refers to the content-aware semantic modeling, **DAB** refers to the dual adaptive branches including **CFB** and **RRB**. We observe that

CSM	DAB		RVOS→A2D		UNC→UNC+	
	CFB	RRB	OIoU	mAP*	OIoU	mAP*
			40.21	21.36	34.57	23.15
✓			40.16	21.44	34.82	23.46
✓	✓		44.91	26.10	36.73	25.33
✓		✓	44.26	25.30	36.09	24.87
	✓	✓	41.92	23.07	35.12	24.23
✓	✓	✓	**46.31**	**27.32**	**37.89**	**26.34**

Table 3: Main ablation study on two transfer tasks. OIoU=Overall IoU. mAP*=mAP[0.5, 0.95].

Ablation Method		RVOS→A2D		UNC→UNC+	
		OIoU	mAP*	OIoU	mAP*
AE	w/ mean-pooling(v)	41.42	22.60	34.21	22.76
	w/ mean-pooling(t)	43.92	24.86	36.44	25.26
	w/o predictor	45.12	26.25	37.24	25.88
SC	w/o boundary	45.34	26.36	37.17	25.79
SI	w/o $\mathcal{L}_{si}$	44.76	25.63	36.13	24.96
	Full	**46.31**	**27.32**	**37.89**	**26.34**

Table 4: Ablation results about the CSM module.

Ablation Method		RVOS→A2D		UNC→UNC+	
		OIoU	mAP*	OIoU	mAP*
CFB	w/o contrastive	44.96	25.83	36.28	24.92
	w/o memory	45.24	26.58	37.21	25.83
RRB	w/o $\mathcal{L}_{ct}$	45.40	26.62	37.32	25.89
	w/o $\mathcal{L}_{ho}$	46.02	26.84	37.64	26.12
	w/o reciprocal	45.21	26.49	37.18	25.61
	Full	**46.31**	**27.32**	**37.89**	**26.34**

Table 5: Ablation results about the DAB module.

only adding the CSM module brings little gains, which is reasonable since it mainly models the content shift for subsequent adaptation. On the basis of CSM, the CFB and RRB both improve the performance dramatically, verifying their effectiveness to address the feature- and relation-level discrepancies. To evaluate the importance of CSM, we remove it and obtain inferior results, confirming the necessity of semantic modeling to harmonize the adaptation. Our full model integrates these modules and therefore achieves better performance.

Ablation Study for the CSM Module. We perform ablation study for the CSM module and report the results in Section 4.3. (1) For attentive extraction (**AE**), we adopt mean-pooling on the visual and textual features as content features to generate ablation models **w/ mean-pooling(v)** and **w/ mean-pooling(t)** respectively. By comparison, our attentive extraction leads to superior performance, evidencing our discussion in Section 3.2. We also remove the predictor as **w/o predictor** and observe it can bring improvement. (2) For semantic construction (**SC**), we remove the boundary for filtering as **w/o boundary**. The result indicates that filtering the unreliable samples is essential for adaptation. (3) For semantic initialization (**SI**), we remove the initialization as **w/o $\mathcal{L}_{si}$**. We also draw the performance curve of stage 3 (Section 3.4) in Figure 4 to reflect the adaptation. The results show its necessity for the superior and stable performance.

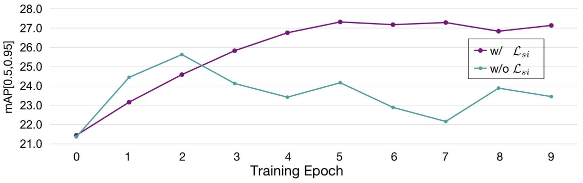
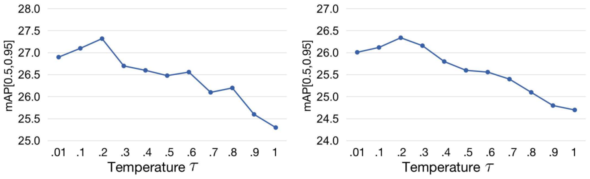
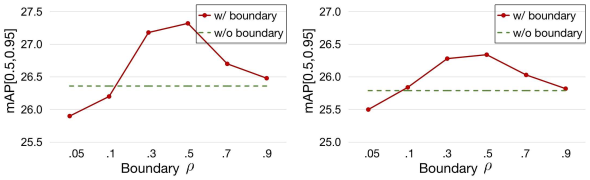


Figure 4: Effect of Loss $\mathcal{L}_{si}$ on RVOS→A2D.



(a) RVOS→A2D (b) UNC→UNC+

Figure 5: Impact of Temperature τ on two transfer tasks.



(a) RVOS→A2D (b) UNC→UNC+

Figure 6: Impact of Boundary ρ on two transfer tasks.

Ablation Study for the DAB Module. We next perform ablation study for the DAB module and report the results in Section 4.3. (1) For the contrastive feature branch (**CFB**), we discard the contrastive learning and directly enhance the pixel similarity as **w/o contrastive**. It is observed that the contrastive learning can effectively promote feature adaptation. We also remove the memory bank as **w/o memory**, and find the memory mechanism can bring further gains. (2) For the reciprocal relation branch (**RRB**), we discard the collaborative training (CT) and hierarchical optimization (HO) as **w/o $\mathcal{L}_{ct}$** and **w/o $\mathcal{L}_{ho}$**, respectively. As illustrated, both components contributes to better adaptation.

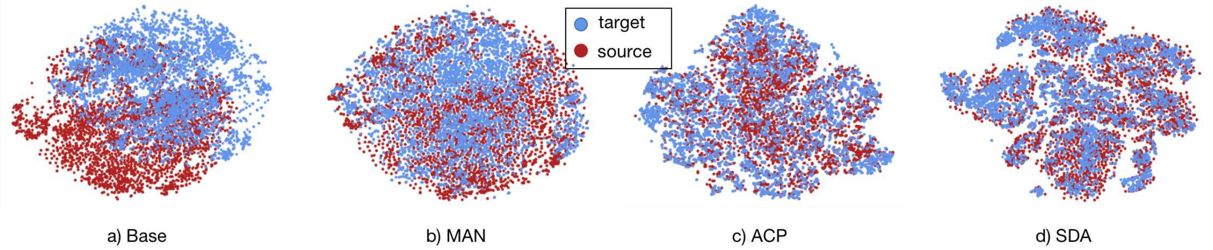


Figure 7: The t-SNE visualization of the visual features on RVOS→A2D.

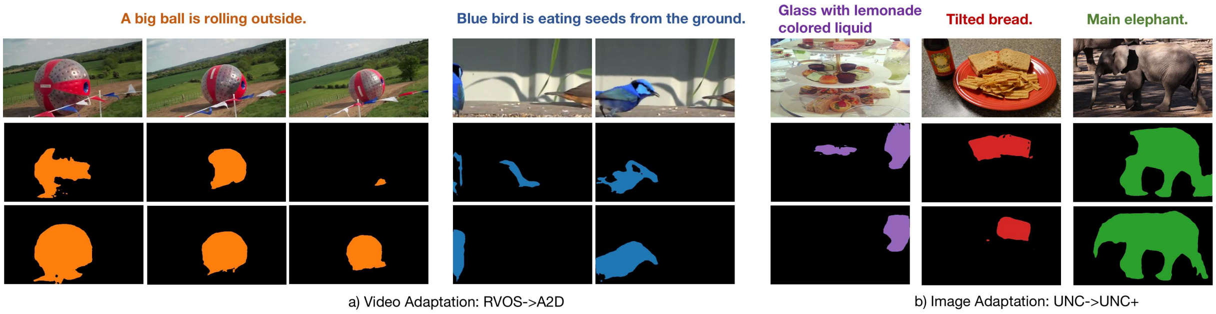

A big ball is rolling outside.
Blue bird is eating seeds from the ground.
Glass with lemonade colored liquid
Tilted bread.
Main elephant.
a) Video Adaptation: RVOS->A2D
b) Image Adaptation: UNC->UNC+

Figure 8: The segmentation results on RVOS→A2D and UNC→UNC+. Base shown in the second row. SDA shown in the third row.

We further replace the reciprocal masks with only domain-based masks as **w/o reciprocal**. The inferior results confirm its effectiveness for relation enhancement.

4.4 Hyper-Parameter Analysis

Impact of Temperature τ in AE method. Temperature τ controls the attention distribution for the extraction of content features. We evaluate 11 different τ values from 0.01 to 1.0 on RVOS →A2D and UNC→UNC+. The result in Figure 5 shows that the performance achieves the best when τ is set to 0.2 and becomes poor when τ is too small or too large. This result suggests that a proper τ value is crucial to capturing key contents.

Impact of Boundary ρ in SC method. To study the impact of boundary ρ, we set $\rho = \beta \log(k^s)$ where $\log(k^s)$ is the theoretical maximum value of the similarity entropy $\mathcal{H}(d_i)$. The result in Figure 6 shows that the performance increases and then decreases with increasing β, indicating that the boundary controls the openness degree between domains and hence affects adaptation.

4.5 Qualitative Analysis

To shed a qualitative light on evaluating the proposed approach, we conduct several experiments as follows. More results are listed in Appendix D.3

Visualization of Visual Features. In Figure 7, we visualize the visual features on RVOS→ A2D, learned by Base, MAN, ACP and SDA respectively using t-SNE (Donahue et al., 2014). The visual features learned by our SDA obtain a clearer boundary and a better separation, indicating the cross-domain visual features are aligned in a category-wise manner under the common semantic guidance.

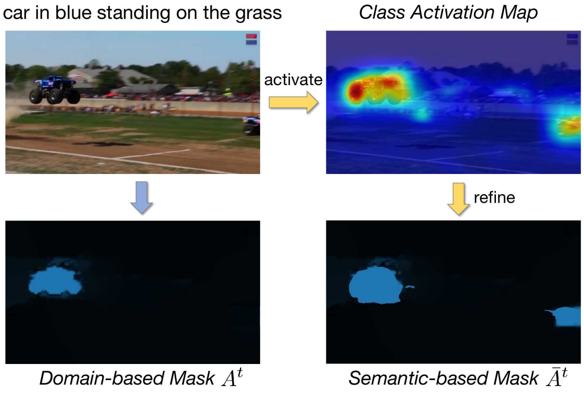

car in blue standing on the grass
Class Activation Map
activate
refine
Domain-based Mask A^t
Semantic-based Mask $\bar{A}^t$

Figure 9: The visualization of reciprocal masks.

Visualization of Segmentation Result. As shown in Figure 8, we visualize the segmentation results to verify the effectiveness of our method. Notice that the results from the Base model are more inaccurate and biased due to the large domain gap, while our SDA produces more accurate segmentation results.

Visualization of Reciprocal Masks. In Figure 9, we visualize the reciprocal masks. The domain-based mask can accurately segment an instance (i.e.

the left car) but suffers severe bias. Instead, the semantic-based mask can coarsely localize various car instances, thus providing comprehensive clue about the missing one (i.e. the right car).

5 Conclusion

In this work, we first study the task of cross-domain query-based visual segmentation. To address this problem, we propose Semantic-conditioned Dual Adaptation, a novel framework that achieves the feature- and relation-level adaptation via a universal semantic structure. Experiments shows that our framework performs consistently well on both query-based video and image benchmarks.

6 Limitation

In this section, we make a clear discussion of the limitation of our work. Our work mainly study the setting where each dataset serves as an independent domain. However, the adopted datasets (e.g. UNC, UNC+) for query-based image segmentation are mostly collected on MS-COCO (Lin et al., 2014) and have limited domain gap between visual modality. The findings could inspire the researchers to explore other settings, e.g. each class serves as an independent domain.

7 Ethics Statement

We adopt the widely-used datasets that were produced by previous researchers. We followed all relevant legal and ethical guidelines for their acquisition and use. Besides, we recognize the potential influence of our technique, such as its application in human-computer interaction and vision-language grounding system. We are committed to conducting our research ethically and ensuring that our research is beneficial. We hope our work can inspire more investigations for the domain adaptation on multi-modal tasks and wish our framework can serve as a solid baseline for further researches.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No.62222211, Grant No.61836002 and Grant No.62072397, and Yiwise.

References

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282.

Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. 2014. Domain adaptation on the statistical manifold. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2488.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. 2022. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995.

Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. 2022. Localized vision-language matching for open-vocabulary object detection. *arXiv preprint arXiv:2205.06160*.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. 2019a. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 627–636.

Qingchao Chen, Yang Liu, and Samuel Albanie. 2021. Mind-the-gap! unsupervised domain adaptation for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1072–1080.

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE international conference on computer vision*, pages 521–530.

Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. 2019b. Referring expression object segmentation with caption-aware consistency. *arXiv preprint arXiv:1910.04748*.

Xize Cheng, Linjun Li, Tao Jin, Rongjie Huang, Wang Lin, Zehan Wang, Huangdai Liu, Ye Wang, Aoxiong Yin, and Zhou Zhao. 2023. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. *arXiv preprint arXiv:2303.05309*.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer.

Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. 2020. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10488–10497.

Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, and Fei Wang. 2021. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196.

Kanishk Jain and Vineet Gandhi. 2021. Comprehensive multi-modal interactions for referring image segmentation. *arXiv preprint arXiv:2104.10412*.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. 2013. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199.

Tao Jin and Zhou Zhao. 2021a. Contrastive disentangled meta-learning for signer-independent sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5065–5073.

Tao Jin and Zhou Zhao. 2021b. Generalizable multilinear attention network. *Advances in Neural Information Processing Systems*, 34:9049–9060.

Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022. Mc-slt: Towards low-resource signer-adaptive sign language translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4939–4947.

Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Cross-modal cross-domain moment alignment network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10678–10686.

Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. 2020. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 33:3569–3580.

Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. 2022. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. 2017. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1271–1280.

Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. 2021a. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang Liu, Qingchao Chen, and Samuel Albanie. 2021b. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14954–14964.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR.

Wenfeng Luo and Meng Yang. 2020. Semi-supervised semantic segmentation via strong-weak dual-branch network. In *European Conference on Computer Vision*, pages 784–800. Springer.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645.

Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2020. Visual-textual capsule routing for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9942–9951.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, and Tao Mei. 2020. Exploring category-agnostic clusters for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13867–13875.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shuang Qiu, Yao Zhao, Jianbo Jiao, Yunchao Wei, and Shikui Wei. 2019. Referring image segmentation by generative adversarial learning. *IEEE Transactions on Multimedia*, 22(5):1333–1344.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pages 208–223. Springer.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Alice Thomas and Glenda Thorne. 2009. How to increase higher order thinking. *Metarie, LA: Center for Development and Learning*, page 264.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526.

Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. 2020. Context modulated dynamic networks for actor and action video segmentation with language queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12152–12159.

Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. 2019. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3939–3948.

Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. 2015. Can humans fly? action understanding with multiple classes of actors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2264–2273.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. 2018. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119.

Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2021. Simulslt: End-to-end simultaneous sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4118–4127.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424.

Wei Zhang, Xiaogang Wang, Deli Zhao, and Xiaoou Tang. 2012. Graph degree linkage: Agglomerative clustering on a directed graph. In *European conference on computer vision*, pages 428–441. Springer.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

This appendix contains four sections. (1) Appendix A introduces the detailed design of our base segmentation network. (2) Appendix B introduces the technique components in our SDA framework, including the predictor (Appendix B.1) and the multi-label visual classification (Appendix B.2). (3) Appendix C introduces the experiment details, including the dataset details (Appendix C.1), the implementation details (Appendix C.2), and the baseline settings (Appendix C.3). (4) Appendix D presents extensive experiment results, including some discussions (Appendix D.1), more hyper-parameter analysis (Appendix D.2) and more qualitative results (Appendix D.3).

A Base Segmentation Network

We adopt a unified segmentation network for both query-based video and image segmentation. Specifically, we adopt the same architecture including the query encoder, the interaction mechanism between the frames and words, the decoder for segmentation and the training loss. The main difference is that we adopt different visual encoders for videos and images, respectively.

Encoder. For each video, we employ pre-trained I3D layers (Carreira and Zisserman, 2017) with stacked 3D convolution to learn the spatio-temporal features for video clips, denoted as $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times C}$, where T, H, W, C are the frame number, height, width and channel number of output respectively. For each image, we employ a pre-trained ResNet-101 network (He et al., 2016) to learn the spatio features, denoted as $\mathbf{V} \in \mathbb{R}^{H\times W\times C}$. For ease of presentation, we abuse the symbol $\mathbf{V}$ to denote both video and image features and drop the T. Besides, we adopt the multi-resolution visual features maps $\{\mathbf{V}_i \in \mathbb{R}^{H_i\times W_i\times C_i}\}_{i=1}^{N_m}$ that are outputs of different encoder layers, where N_m is the number of multi-resolution feature maps, H_i, W_i and C_i are separately the width, height and channel number of the i-th feature map. For each query, we employ the Glove (Pennington et al., 2014) word embeddings as the input $\mathbf{W}$ and apply a Bi-GRU network to learn the query features $\mathbf{Q} \in \mathbb{R}^{N\times C}$, where N is the word number.

Interaction. With the query representation $\mathbf{Q} \in \mathbb{R}^{N\times C}$ and the visual features $\mathbf{V} \in \mathbb{R}^{H\times W\times C}$, we first incorporate the natural language to generate query-focused visual context $\mathbf{V}_q \in \mathbb{R}^{H\times W\times C}$ through a dot-product attention and gating modulation, given by:

$$\begin{aligned}\bar{\mathbf{V}} &= \mathrm{softmax}(g_1(\mathbf{V})g_2(\mathbf{Q}^\top))\mathbf{Q} \\ \mathbf{V}_q &= tanh(g_3(\bar{\mathbf{V}}))\mathbf{V}\end{aligned} \tag{8}$$

where g_1, g_2, g_3 are distinct linear transformations and $\bar{\mathbf{V}} \in \mathbb{R}^{H\times W\times C}$ is the attentive representation. Note that the cross-modal attention α in Section 3.1 can be obtained by $\alpha = \mathrm{softmax}(g_1(\mathbf{V})g_2(\mathbf{Q}^\top))$. We then divide the feature map into different semantic regions based on the unsupervised low-level SLIC superpixel algorithm (Achanta et al., 2012). Specifically, we turn the visual context $\mathbf{V}_q$ into the superpixel representation $\mathbf{V}_r \in \mathbb{R}^{N_r\times C}$ through region max-pooling, where N_r is the preset number of superpixels, and compute the region-contextual representations $\bar{\mathbf{V}}_r$ with region self-attention, given by:

$$\bar{\mathbf{V}}_r = \mathrm{softmax}(g_q(\mathbf{V}_r)g_k(\mathbf{V}_r^\top))g_v\mathbf{V}_r \tag{9}$$

where g_q, g_k, g_v are distinct linear transformation. We further augment the pixel representations by adding corresponding region-contextual representations to the original visual context and get the enhanced pixel-level features $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$.

We build the hierarchical cross-modal interaction to obtain enhanced feature maps $\{\mathbf{F}_i \in \mathbb{R}^{H_i\times W_i\times C_i}\}_{i=1}^{N_m}$ by stacking the interaction module over multi-resolution visual features $\{\mathbf{V}_i\}_{i=1}^{N_m}$.

Decoder. After the encoding and interaction, we generate the multi-scale response maps $\{\{\mathbf{s}_{i,j}\}_{j=1}^{H_i\times W_i}\}_{i=1}^{N_m}$ by employing FCN (fully convolutional network) on the enhanced representation $\{\mathbf{F}_i\}_{i=1}^{N_m}$.

Training. With the multi-scale response maps $\{\{\mathbf{s}_{i,j}\}_{j=1}^{H_i\times W_i}\}_{i=1}^{N_m}$ and pixel-wise annotations $\{\{a_{i,j}\}_{j=1}^{H_i\times W_i}\}_{i=1}^{N_m}$ where $a_{i,j} \in \{0, 1\}$, we directly compute the binary cross-entropy loss for j-th pixel of i-th feature map, given by

$$\begin{aligned}\mathcal{L}_{seg,(i,j)} = &- a_{i,j}\mathrm{log}\mathbf{s}_{i,j} \\ &- (1 - a_{i,j})\mathrm{log}(1 - \mathbf{s}_{i,j})\end{aligned} \tag{10}$$

B Technique Components

B.1 Predictor

In the CSM module (Section 3.2), we independently train a predictor to predict the weight for the words on the target domain, aiming to directly learn the importance from the query without available annotations. The predictor consists of a two-layer MLP. Specifically, we conduct pre-training on the source domain where we fix all other network components except for the predictor. First, we input the query embeddings $\mathbf{W}$ into the predictor and obtain the output $\{\bar{\alpha}_n^{s,pre}\}_{n=1}^N$. Then we follow the knowledge distillation scheme (Hinton et al., 2015) to adopt the visual-guided attention weights $\{\bar{\alpha}_n^s\}_{n=1}^N$ as the objective and develop the L1 loss to train the predictor. The loss is given by:

$$\mathcal{L}_{pre} = D_{\mathrm{KL}}(\bar{\alpha}_n^{s,pre}||\bar{\alpha}_n^s) \tag{11}$$

where $D_{\mathrm{KL}}(A||B)$ is the Kullback-Leibler divergence from A to B. After training the predictor for 5 epochs, we freeze it and apply it to the target domain to predict the weight $\bar{\alpha}_n^t$. The visualized results can be found in Figure 12.

B.2 Multi-label Visual Classification

In the RRB branch (Section 3.3.2), with the semantic category L^t for each referred instance, we first construct image-level training samples by combining all categories that appeared in the image. In this way, an image (or a video frame) is related to multiple category labels and we assume that the ground truth label of an image is $y \in \mathbb{R}^{k^t+k^u}$, where $y^i = \{0, 1\}$ denotes whether label i appears in the image or not. Next, we leverage an independent classification network to perform training, where we adopt a pre-trained ResNet-101 as the backbone to encode visual features. Then we apply the global average pooling on the convolutional feature maps and input them into a fully-connected layer with a sigmoid function to produce the desired output $\bar{y}$ for classification. With total $k^t + k^u$ categories on the target domain, the whole network is trained using the traditional multi-label classification loss as follows:

$$\mathcal{L}_{mlc} = -\sum_{c=1}^{k^t+k^u} (y^c \log \bar{y}^c + (1-y^c)\log(1-\bar{y}^c)) \tag{12}$$

To leverage the weakly-supervised localization ability, we follow the CAM method (Zhou et al., 2016) to build the class activation map, which can coarsely highlight the pixels belonging to a specified category. We also follow the IRN method (Ahn et al., 2019) to further refine the class activation map for more precise masks. Afterwards, each target training sample (V_i^t, Q_i^t) is associated with a semantic-based mask $\bar{A}_i^t$, which provides the coarse-level instance location of the corresponding category L_i^t. The visualized results are shown in Figure 9.

C Experiment Details

C.1 Dataset Details

C.1.1 Query-based Video Segmentation

Refer-Youtube-VOS. Refer-Youtube-VOS (Seo et al., 2020) is a large-scale referring video segmentation dataset extended from Youtube-VOS dataset (Xu et al., 2018) which contains 3975 videos, 7451 objects and 27899 expressions with both first-frame expression and full-video expression annotated.

A2D Sentences. A2D Sentences (Gavrilyuk et al., 2018) is extended from the Actor-Action Dataset (Xu et al., 2015) by providing textual descriptions for each video. It contains 3,782 videos annotated with 8 action classes performed by 7 actor classes.

J-HMDB Sentences. J-HMDB sentences (Gavrilyuk et al., 2018) is extended from the J-HMDB dataset (Jhuang et al., 2013) which contains 928 videos and corresponding 928 sentences. All the actors in JHMDB dataset are humans and one natural language query is annotated to describe the action performed by each actor.

C.1.2 Query-based Image Segmentation

UNC. UNC (Yu et al., 2016) is collected on MS-COCO (Lin et al., 2014). It contains 19, 994 images with 142,209 referring expressions for 50,000 objects. Expressions in UNC contain words indicating the location of the objects.

UNC+. UNC+ (Yu et al., 2016) is also collected on MS-COCO (Lin et al., 2014). It contains 19, 992 images with 141,564 referring expressions for 49,856 objects. Expressions in UNC+ describe the objects based on their appearance and context within the scene without using spatial words.

G-Ref. G-Ref (Mao et al., 2016) is also collected on MS-COCO (Lin et al., 2014). It contains 26,711 images with 104,560 referring expressions. Expressions in G-Ref contain longer sentences with an average length of 8.4 words compared with other datasets (e.g. UNC, UNC+) which have an average sentence length of less than 4 words.

C.2 Implementation Details

Model Selection. For visual features, we use the ResNet-101 (He et al., 2016) pre-trained on the ImageNet as the backbone feature extractor for images and use the I3D network (Carreira and Zisserman, 2017) pre-trained on the Kinetics dataset (Carreira et al., 2018) for video clips. For query features, we employ the pre-trained Glove (Pennington et al., 2014) word embeddings as input.

Parameter Setting. For the base segmentation setting, we follow the video segmentation work (Wang et al., 2019) to set the target frame as the center of 8 continuous clips. All the frames are rescaled and padded to the same size of 320 × 320. The FCN network for the decoder consists of three fully convolutional layers with residual connection, where the kernel size is 3 × 3 for the first two layers and 1 × 1 for the remaining layer. We set the hidden size to 1024 and use bilinear interpolation for feature map upsampling. For images, we adopt the last three layers of the encoder for the multi-resolution feature maps ($N_m = 3$). For videos, we adopt the last five layers of the encoder for the multi-resolution feature maps ($N_m = 5$).

In our SDA framework, we select the visual features from the last layer of the encoder for adaptation. For the content-aware semantic modeling module, we set the temperature τ to 0.2, set the distance threshold in Agglomerative Clustering (Zhang et al., 2012) to 0.5 and set the boundary ρ to $\frac{\log(k^s)}{2}$ where $\log(k^s)$ is the theoretical maximum value of entropy $\mathcal{H}(d)$. For the contrastive feature branch, we set the thresholds γ_{max} and γ_{min} to 0.9 and 0.1 respectively and set the memory size B to 100 for each category. Besides, we adopt the teacher-student architecture (Tarvainen and Valpola, 2017) to provide stable features with the momentum parameter set to 0.99. For the reciprocal feature branch, we follow the IRN method (Ahn et al., 2019) to refine the class activation map. The loss coefficients λ_1 and λ_2 are empirically fixed at 1.0 and 0.1. To train our model, we use the Adam optimizer with an initial learning rate 1e-7. The learning rate increases to 4e-4 linearly for 300 updating steps and then decreases proportionally. The batch size is set to 8 for both the source data and target data. We run all the experiments for 5 times and report the mean results.

Training Step. As mentioned in Section 3.4, we develop a multi-stage training. In stage 1, we pre-train the segmentation model with BCE loss $\mathcal{L}_{seg}$ on the source domain for 20 epochs. In stage 2, we re-train the model by adding the loss $\mathcal{L}_{si}$ in CSM module for 20 epochs. In stage 3, we train the model with the full loss $\mathcal{L}_{sda}$ for 10 epochs. More specifically, in stages 1 and 2, we both set the learning rate to 4e-4 to start training. In stage 3, we continue training with the updated learning rate.

Experiment Configuration. The SDA is implemented using PyTorch 1.9.0 with CUDA 10.0 and cudnn 7.6.5. All the experiments are conducted on a workstation with four NVIDIA GeForce RTX 2080Ti GPUs.

C.3 Baseline Setting

Query-based Visual Segmentation Baselines. For video segmentation baselines, **CMDy-Conv** (Wang et al., 2020) utilizes a context modulated dynamic network with group-wise kernel prediction to incorporate context information and an effective temporal evolution encoder to capture motion information; **CMPC-V** (Hui et al., 2021) builds a cross-modal adaptive modulation module to dynamically recombine multi-modal features. For image segmentation baselines, **CSMA** (Ye et al., 2019) employs cross-modal attention and self-attention to extract multi-modal context between image regions and referring words; **CMPC-I** (Hui et al., 2021) applies the same architecture as CMPC-V without temporal interaction on the image side. We directly re-implement the above approaches and adopt the same visual encoder and query embedding for a fair comparison.

Domain Adaptation Baselines. We combine domain adaptation approaches with our Base segmentation network to conduct experiments. For DA baselines designed for uni-modal tasks, **MMD** (Long et al., 2015) minimizes the feature distances; **DANN** (Ganin and Lempitsky, 2015) employs a gradient reversal layer to learn domain-invariant features; **PLCA** (Ganin and Lempitsky, 2015) develops pixel-level contrastive learning based on the pixel similarity. To apply MMD and DANN on the CQVS task, we leverage them for both pixel-level visual features and query features. Since PLCA mainly works for visual pixels, we further apply MMD on query features to improve the performance. For DA baselines designed for multi-modal tasks, **MAN** (Jing et al., 2020) performs alignment for each modality feature and leverages pseudo labels to train the target samples; **ACP** (Liu et al., 2021a) employs the pre-trained classification model to preserve the semantic structure of compositional concepts from uni-modal data. Since both MAN and ACP are designed for global image-level features, we apply MAN by replacing the image-level pseudo labels with pixel-level pseudo labels, and apply ACP by replacing the image-level con-

Algorithm	Cluster	RVOS→A2D	
	Number	OIoU	mAP*
K-Means	10	43.42	24.20
	50	44.64	25.86
	100	44.97	26.05
	500	44.11	24.63
Spectral	10	43.38	24.26
	50	45.14	25.93
	100	45.42	26.33
	500	44.20	24.48
Agglomerative	-	**46.31**	**27.32**

Table 6: The Comparison of different clustering algorithms.

Feature	RVOS→A2D		UNC→UNC+	
	OIoU	mAP*	OIoU	mAP*
Encoded Features	45.36	26.48	37.07	25.39
Word Embedding	**46.31**	**27.32**	**37.89**	**26.34**

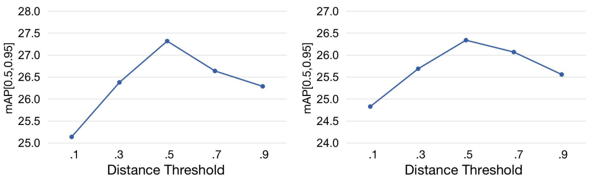
Table 7: The Comparison of Content Extraction: Encoded Features vs Word Embedding.

Size	10	50	100	200
mAP*	26.83	27.14	27.32	**27.36**

Table 8: Impact of Size B on the RVOS → A2D task.

γ_{max}	γ_{min}	RVOS→A2D	
		oIoU	mAP*
0.9	0.1	**46.31**	**27.32**
0.8	0.2	45.52	26.54
0.7	0.3	44.96	26.11
0.9	0.2	45.73	26.85
0.9	0.3	45.24	26.22
0.8	0.1	46.17	27.09
0.7	0.1	45.82	26.94

Table 9: Impact of Thresholds γ_{max} and γ_{min}.



(a) RVOS→A2D (b) UNC→UNC+

Figure 10: Impact of Distance Threshold on two transfer tasks.

cept with the instance-level concept (i.e. average pooling of pixels in the foreground region).

D Experiment Results

D.1 Discussion

Selection of Clustering Algorithm. To effectively establish the semantic structure of multi-modal contents, we investigate different clustering algorithms including K-Means (Lloyd, 1982), Spectral Clustering (Ng et al., 2001) and Agglomerative Clustering (Zhang et al., 2012). From the results shown in Appendix D, we observe our framework is sensitive to the choice of a specific clustering algorithm. Specifically, K-Means and Spectral Clustering both require a cluster number value that is manually set. The optimal cluster number is difficult to define, hindering these algorithms to obtain satisfactory clustering results and leading to inferior adaptation performance. Instead, Agglomerative Clustering only requires a proper distance threshold parameter to automatically perform hierarchical clustering by grouping similar points, which yields better results on our content features extracted from the embedding model.

Content Extraction: Encoded Features vs Word Embedding. As discussed in Section 3.2, the work (Bravo et al., 2022) shows that a simple language model fits better than a large contextualized language model for detecting novel objects. To further verify its effectiveness on semantic modeling, we compare two different query features for content extraction, i.e. the encoded features vs word embedding. We present the results in Appendix D and find that the word embedding performs better than encoded features, which is consistent with the conclusion (Bravo et al., 2022).

D.2 Hyper-Parameter Analysis

Impact of Memory Size B in CFB module. We set the memory size B to [10, 50, 100, 200] to explore the impact of it. The result in Table 8 reveals that a larger memory size can bring more improvement, which is also verified in (He et al., 2020). Notably, when the size B exceeds 100, the gain is limited. Considering the computation cost, we set $B = 100$ in our experiments.

Impact of Thresholds γ_{max} and γ_{min} in the CFB module. Thresholds γ_{max} and γ_{min} separately control the number of selected pseudo pixels for foreground and background in the target domain. We analyze the impact of two thresholds and report the results in Appendix D.2. It indicates that both threshold values are crucial to the adaptation performance. Our SDA achieves the best results when the values γ_{max}=0.9 and γ_{min}=0.1. It also shows

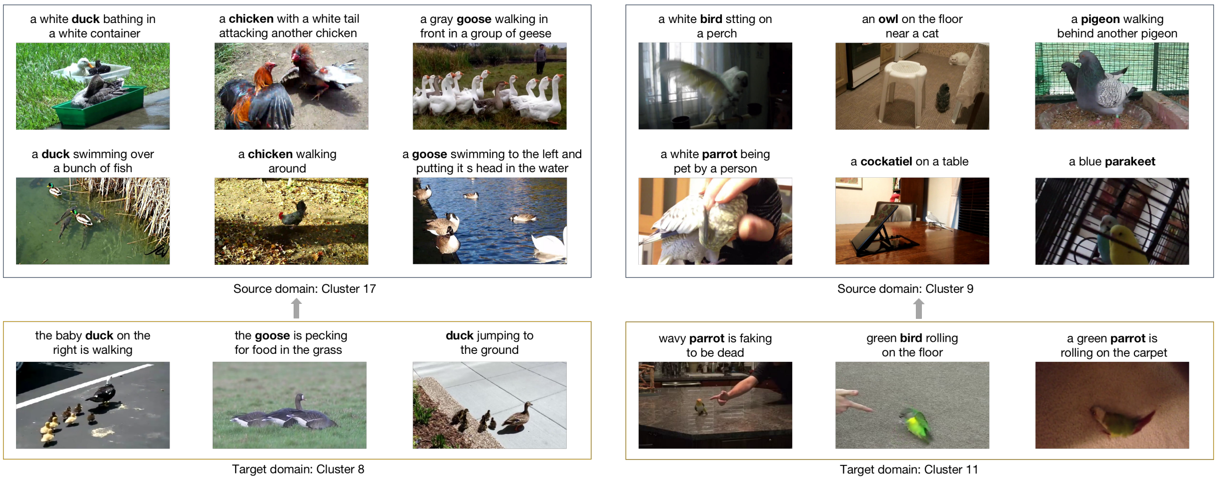
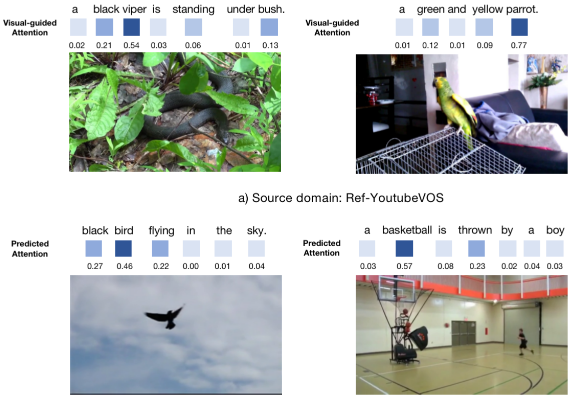

a white **duck** bathing in a white container
a **chicken** with a white tail attacking another chicken
a gray **goose** walking in front in a group of geese
a **duck** swimming over a bunch of fish
a **chicken** walking around
a **goose** swimming to the left and putting it s head in the water
Source domain: Cluster 17
the baby **duck** on the right is walking
the **goose** is pecking for food in the grass
duck jumping to the ground
Target domain: Cluster 8
a white **bird** stting on a perch
an **owl** on the floor near a cat
a **pigeon** walking behind another pigeon
a white **parrot** being pet by a person
a **cockatiel** on a table
a blue **parakeet**
Source domain: Cluster 9
wavy **parrot** is faking to be dead
green **bird** rolling on the floor
a green **parrot** is rolling on the carpet
Target domain: Cluster 11

Figure 11: The semantic clusters on RVOS→A2D.



a) Source domain: Ref-YoutubeVOS

b) Target domain: A2D Sentences

Figure 12: The visualization of visual-guided attention on RVOS→A2D. Darker color means the higher attention score.

that the γ_{min} value is a bit more important than the γ_{max}, since the background pixels can provide the discriminative power serving as negative samples.

Impact of Distance Threshold in the SC method. Distance threshold in Agglomerative Clustering defines the minimum distance between two clusters, which indirectly controls the cluster number. To explore the impact of it, we set the distance threshold to [0.1, 0.3, 0.5, 0.7, 0.9] and display the results in Figure 10. We note that the performance gradually improves with the increase of distance threshold and slowly reaches the bottleneck. This phenomenon is reasonable since a large threshold leads to few clusters where each one contains many indistinguishable samples and a small threshold results in too many clusters where each one has few samples.

D.3 Qualitative Analysis

Visualization of Semantic Construction. In Figure 11, we present the semantic clusters on each domain. We observe that the cluster on the source domain can group semantically similar visual instances and sentences, e.g. "bird", "parrot" and "owl". Meanwhile, the separation between clusters is also clear, e.g. "bird" vs "duck". Thus, the similar parts on the target domain can be well-aligned to the source clusters. The visualization results demonstrate the effectiveness of our content-aware semantic modeling module to explore the multi-modal contents and learn the semantic structure across domains.

Visualization of Attentive Extraction. In Figure 12, we depict the distribution of the visual-guided attention on the source domain and the predicted attention on the target domain. We can find the attention weights on both domains can highlight the crucial words in the query, e.g. the actor "viper" and "bird", while suppressing the inessential parts, e.g. the descriptive words "under bush" and "in the sky".

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6.

☑ A2. Did you discuss any potential risks of your work?
Section 7.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4: Experiment.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C.2.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C.2.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix C.2.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix C.2.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.