Scene-robust Natural Language Video Localization via Learning Domain-invariant Representations

Zehan Wang and **Yang Zhao** and **Haifeng Huang** and **Yan Xia** and **Zhou Zhao**[*]
{wangzehan01,awalk,huanghaifeng,zhaozhou} @zju.edu.cn
Zhejiang University

Abstract

Natural language video localization(NLVL) task involves the semantic matching of a text query with a moment from an untrimmed video. Previous methods primarily focus on improving performance with the assumption of independently identical data distribution while ignoring the out-of-distribution data. Therefore, these approaches often fail when handling the videos and queries in novel scenes, which is inevitable in real-world scenarios. In this paper, we, for the first time, formulate the scene-robust NLVL problem and propose a novel generalizable NLVL framework utilizing data in multiple available scenes to learn a robust model. Specifically, our model learns a group of generalizable domain-invariant representations by alignment and decomposition. First, we propose a comprehensive intra- and inter-sample distance metric for complex multi-modal feature space, and an asymmetric multi-modal alignment loss for different information densities of text and vision. Further, to alleviate the conflict between domain-invariant features for generalization and domain-specific information for reasoning, we introduce domain-specific and domain-agnostic predictors to decompose and refine the learned features by dynamically adjusting the weights of samples. Based on the original video tags, we conduct extensive experiments on three NLVL datasets with different-grained scene shifts to show the effectiveness of our proposed methods.

1 Introduction

Natural language video localization(NLVL) (Wang et al., 2020; Zhao et al., 2021; Zhang et al., 2021) aims to retrieve a temporal moment from an untrimmed video to match a language query semantically. As a fundamental vision-language problem, NLVL attracts increasing attention, and recent works (Chen et al., 2018; Gao et al., 2017; Xu et al., 2019a, 2018; Zhang et al., 2020a,b; Chen et al., 2019; Ghosh et al., 2019; Yuan et al., 2019b) has achieved impressive results. However, their performances are tested with the assumption of independently identical data distribution. Given the variety of video content, it is impossible to cover all scenes when building training datasets. Therefore, these methods are impractical in real-world scenarios since they ignore the generalization ability to novel scenes. In this paper, we, for the first time, propose the scene-robust NLVL problem.

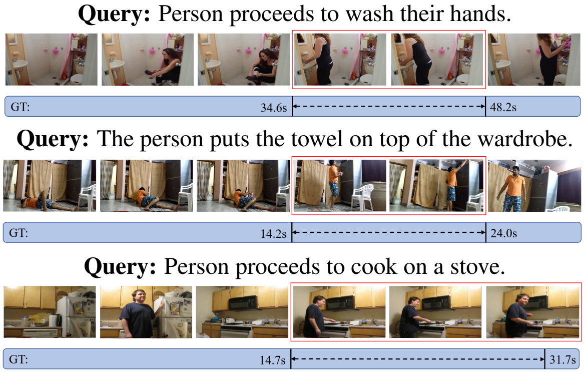

Query: Person proceeds to wash their hands.
GT:
34.6s
48.2s
Query: The person puts the towel on top of the wardrobe.
GT:
14.2s
24.0s
Query: Person proceeds to cook on a stove.
GT:
14.7s
31.7s

Figure 1: An illustration of different scenes in an NLVL dataset. These examples are selected from Charades-STA, and the videos are in *Bathroom*, *Living room* and *Kitchen*, respectively.

Scene-robust NLVL aims to take advantage of data in available scenes to learn a robust model which can generalize well on novel scenes. As shown in the Figure 1, there are obvious semantic differences between video and text input scenes, such as video background, specific actions and objects in a scene. In more detail, we analyze the semantic distribution gaps across scenes in three NLVL datasets. The detailed statistics and discussions in the appendix show that the domain gaps between scenes are prevalent and diverse. Traditional NLVL models would experience dramatic performance degradation when dealing with such semantic distribution shifts.

The proposed scene-robust NLVL problem is

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 144–160
July 9-14, 2023 ©2023 Association for Computational Linguistics

more challenging than the relevant traditional domain generalization(DG) problem. On the one hand, inputs of NLVL are multi-modal, which means the features used for predicting are a fusion of visual and textual features, and the final feature space is much more complex. Besides, the different information attribute of video and text brings more difficulty. Recent DG methods mainly focus on learning domain-invariant representations on single-modal tasks, which are not comprehensive and appropriate for multi-modal features. On the other hand, NLVL requires more reasoning steps with diverse modules, and some samples may mainly rely on domain-specific information for reasoning. The knowledge in these data is hard-to-transfer and may affect the learning of domain-invariant representations and the predictor.

To alleviate these challenges, we proposed a novel generalizable NLVL framework to learn stable and high-quality multi-modal domain-invariant representations from alignment and decomposition perspectives. Specifically, we design a multi-modal domain alignment module that contains: an intra- and inter-sample distance metric for aligning the complex multi-modal feature spaces comprehensively, an asymmetric multi-modal aligning strategy for different information densities of textual and visual features, and Dirichlet-Mixup augmentation to compensate for the missing information and increases the domain diversity. Besides, we introduce domain-specific predictors for each domain to refine the learned domain-invariant representations by decomposition, and dynamically suppress the weights of hard-to-transfer samples by simply summing the outputs of domain-specific and domain-agnostic predictors.

Our main contributions can be summarized as follows:

- To the best of our knowledge, we are the first to formulate the scene-robust NLVL problem, which is quite essential for real-world scenarios and fundamentally more challenging.
- We propose a novel generalizable NLVL framework to address the unique challenges of the scene-robust NLVL problem. It learns and refines domain-invariant representations from both alignment and decomposition perspectives.
- The extensive experiments conducted on three NLVL datasets: Charades-STA, ActivityNet-Caption, and YouCook2, demonstrate the effectiveness of our proposed methods.

2 Related work

Natural Language Video Localization. The task of retrieving video segments using language queries is first introduced by (Gao et al., 2017). The previous methods in the field can be categorized into $proposal\text{-}based$ (Gao et al., 2017; Xu et al., 2018, 2019a; Chen et al., 2018; Zhang et al., 2020b; Wang et al., 2020; Zhang et al., 2019; Yuan et al., 2019a) and $proposal\text{-}free$ (Yuan et al., 2019b; Chen et al., 2019; Ghosh et al., 2019; Zhang et al., 2020a; Mun et al., 2020; Rodriguez et al., 2020; Wang et al., 2019) methods. Specifically, $proposal\text{-}based$ methods mainly rely on sliding windows or segment proposal networks to generate proposal candidates from video and retrieve the best matching one for the given semantic information. (Xu et al., 2018) apply the pre-trained segment proposal network (Xu et al., 2017) for proposal candidates generation, (Xu et al., 2019a) further introduce query information to adjust proposals generation process. (Zhang et al., 2020b) first utilizes a 2D temporal map to generate proposals. Considering the redundant computation and low efficiency of the two-stage propose-and-rank, (Yuan et al., 2019b) build a $proposal\text{-}free$ method using attention-based structures to directly regresses temporal locations of the target moment. Inspired by the concept of question answering(Seo et al., 2016; Yu et al., 2018), (Chen et al., 2019; Ghosh et al., 2019) try to formulate NLVL as a span prediction task. (Zhang et al., 2020a) further explains the similarity and differences between NLVL and question answering, and the proposed VSLNet achieves superior performance with simple network architecture.

Domain Generalization. Domain generalization aims to learn a generalizable model from seen source domains to achieve good performance on unseen target domains. Most existing domain generalization methods can be roughly divided into three categories: *representation learning-based* (Li et al., 2018b, 2017; Ghifary et al., 2016; Rahman et al., 2020), *data augmentation-based* (Somavarapu et al., 2020; Zhou et al., 2021; Shu et al., 2021; Mancini et al., 2020; Cheng et al., 2023), and *learning strategy-based* (Mancini et al., 2018; Cha et al., 2021; Finn et al., 2017; Li et al., 2018a).

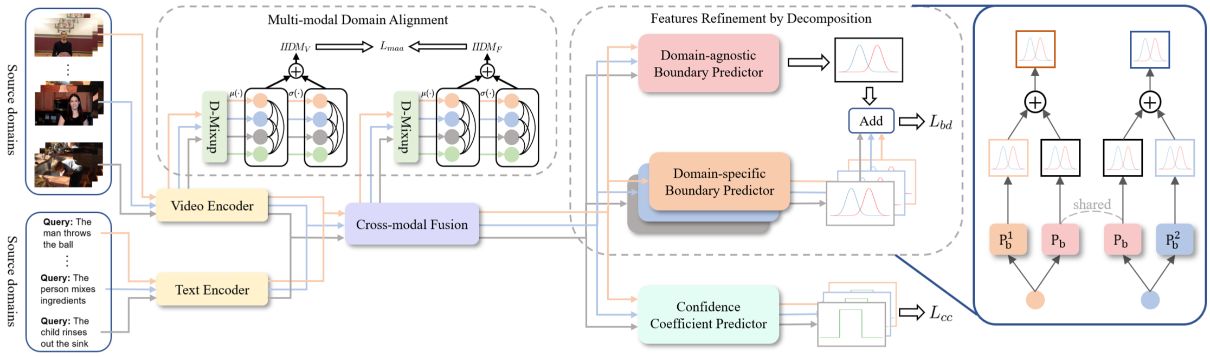

Source domains
Query: The man throws the ball
Query: The person mixes ingredients
Query: The child rinses out the sink
Source domains
Multi-modal Domain Alignment
$IIDM_V$
L_{maa}
$IIDM_F$
D-Mixup
$\mu(\cdot)$
$\sigma(\cdot)$
D-Mixup
$\mu(\cdot)$
$\sigma(\cdot)$
Video Encoder
Text Encoder
Cross-modal Fusion
Features Refinement by Decomposition
Domain-agnostic Boundary Predictor
Add
L_{bd}
Domain-specific Boundary Predictor
Confidence Coefficient Predictor
L_{cc}
P_b^1
P_b
shared
P_b
P_b^2

Figure 2: An overview of our generalizable NLVL framework for scene-robust NLVL problem.

One common approach for domain generalization is learning domain-invariant feature representation, which can be realized by aligning source domains or disentangling features into domain-specific and domain-invariant components. (Li et al., 2018b) employ Adversarial Auto-encoder (Makhzani et al., 2015) and Maximum Mean Discrepancy(Gretton et al., 2012) metric to align distribution between every pair of source domains. (Li et al., 2017) decomposes the model into domain-specific and domain-agnostic parts and only utilizes domain-agnostic parts to make predictions at inference time. *Data augmentation-based* methods focus on manipulating new data to increase the diversity of training data. (Somavarapu et al., 2020) use the style transfer model (Huang and Belongie, 2017) to explore cross-source variability for better generalization. (Zhou et al., 2021; Shu et al., 2021; Mancini et al., 2020) mix the features between instances from different domains to further explore inter-domain feature spaces and compensates for the missing information. *Learning strategy-based* methods employ the general learning strategy to enhance the generalization ability. (Mancini et al., 2018; Cha et al., 2021) exploit ensemble learning to flatten the loss surface and learn a unified and generalized model. Meanwhile, (Li et al., 2018a) proposes meta-learning for domain generalization(MLDG), which uses a meta-learning strategy to achieve domain generalization.

However, most well-studied domain generalization strategies are proposed for single-modal or easy multi-modal tasks requiring less reasoning, such as image classification and image-text retrieval. These tasks' input modality and model structure are much simpler than that of NLVL, and the extra complexity makes the scene-robust NLVL problem more challenging.

3 Approach

We first formulate the scene-robust NLVL problem and our basic model architecture. Then we illustrate our method for learning domain-invariant representation for NLVL from two aspects: multi-modal domain alignment and feature refinement by decomposition.

3.1 Overview

Problem Formulation. Given an untrimmed video $\boldsymbol{v}$ and a related natural language query $\boldsymbol{q}$, NLVL is to retrieve the specific moment $\{\hat{\tau}_s, \hat{\tau}_e\}$ that is most semantically relevant to the given query. Assuming there are K domains in total, for domain l, the input video and query can be denoted as $\boldsymbol{v_l}$ and $\boldsymbol{q_l}$. In scene-robust NLVL, the model can only access several source domains during training, while the target domain's data is unavailable. The generalization ability can be tested on the unseen target domain.

Model Architecture. The overall architecture of our model is illustrated in Figure 2. Concretely, our basic NLVL model consists of a video encoder $e_v : \boldsymbol{v_l} \to \widetilde{\boldsymbol{v_l}} \in \mathbb{R}^{N_v \times D}$, a query encoder $e_q : \boldsymbol{q_l} \to \widetilde{\boldsymbol{q_l}} \in \mathbb{R}^{N_q \times D}$, a cross-modal representation fusion module $m : \widetilde{\boldsymbol{v_l}} \times \widetilde{\boldsymbol{q_l}} \to \boldsymbol{f_l} \in \mathbb{R}^{N_v \times D}$, a domain-agnostic boundary predictor $p_b : \boldsymbol{f_l} \to \{\boldsymbol{T_s}, \boldsymbol{T_e}\}$, and some domain-specific boundary predictors $\{p_b^l\}_{l=1}^K : \boldsymbol{f_l} \to \{\boldsymbol{T_s^l}, \boldsymbol{T_e^l}\}$, where $\boldsymbol{T_s}, \boldsymbol{T_e} \in \mathbb{R}^{N_v}$ denote the scores of start and end boundaries at each frame. Meanwhile, we use a confidence coefficient predictor $p_c : \boldsymbol{f_l} \to \boldsymbol{C} \in \mathbb{R}^{N_v}$ to assist training, where $\boldsymbol{C}$ represents the confidence coefficient of the frame in the matching span.

3.2 Multi-modal Domain Alignment

In this section, we describe our proposed multi-modal domain alignment method. As discussed in (Arora et al., 2017; Long et al., 2018), when the feature distribution is multi-modal, adversarial learning is insufficient to guarantee that the confounded feature distributions are identical, and a simple distance metric would align different domains partially. Therefore, we devise an intra- and inter-sample distance metric to measure the domain differences comprehensively. Considering the multi-modal input in NLVL and the different information densities of each modality, we asymmetrically applied the metric to visual and fusion feature representations to facilitate the learning of multi-modal domain-invariant representations.

Intra- and Inter-sample Distance Metric. For brevity, the latent representations from different domains are denoted as $\{\boldsymbol{H}_l\}_{l=1}^{K}$, where $\boldsymbol{H}_l \in \mathbb{R}^{b \times n \times d}$ and b, n, d represent the number of samples in this batch, feature size and feature dimension of each sample, respectively. First, we calculate the intra- and inter-sample feature distribution:

$$\boldsymbol{H}_{intra}^{l} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{H}_l \tag{1}$$

$$\boldsymbol{H}_{inter}^{l} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{H}_{intra}^{l} - \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{H}_{intra}^{l})^2} \tag{2}$$

The $\boldsymbol{H}_{intra}^{l}$ is the mean value computed across the temporal dimension of each sample, which represents intra-sample feature distribution for source domain l; $\boldsymbol{H}_{inter}^{l}$ is the standard deviation across all samples in a batch, which indicates the inter-sample feature distribution for source domain l.

Based on the complementary two-feature distributions of latent representations, we can comprehensively measure the distance between data distributions of different domains. The proposed intra- and inter-sample distance metric function $IIDM(\cdot)$ is defined as follows:

$$IIDM(\{\boldsymbol{H}_l\}_{l=1}^{K}) = \sum_{1 \leq i,j \leq K} \frac{MMD^2(\boldsymbol{H}_{intra}^{i}, \boldsymbol{H}_{intra}^{j}) + MMD^2(\boldsymbol{H}_{inter}^{i}, \boldsymbol{H}_{inter}^{j})}{K(K-1)} \tag{3}$$

Maximum Mean Discrepancy(MMD) (Gretton et al., 2006) is an effective metric for comparing two distributions. Given feature distributions from domain i and j, the function can be written as:

$$MMD^2(\boldsymbol{H}^i, \boldsymbol{H}^j) = \| \frac{1}{B}\sum_{b=1}^{B}\phi(\boldsymbol{h}_b^i) - \frac{1}{B}\sum_{b=1}^{B}\phi(\boldsymbol{h}_b^j) \|^2 \tag{4}$$

where ϕ is a simple gaussian or linear kernel function that maps representations to a high-dimensional feature space.

Multi-modal Asymmetric Alignment. In the NLVL task, the query sentences are semantically related to the video. For different scenes, there are significant gaps in both visual and textual features. A straightforward approach is to align textual and visual features separately by the proposed distance metric. However, as discussed in (He et al., 2022), the information density is different between language and vision. Textual features are highly semantic and information-dense, while visual features are heavily redundant. Using the same domain alignment method for different modalities would result in either inadequate alignment of visual features or loss of textual semantic information.

To address this dilemma, we devise an asymmetric method to align different modalities. Specifically, the intra- and inter-sample distance metric is applied on the visual features $\{\widetilde{\boldsymbol{v}}_l\}_{l=1}^{K}$ and fused features $\{\boldsymbol{f}_l\}_{l=1}^{K}$, and the total multi-modal asymmetric alignment loss L_{maa} can be formulated as:

$$L_{maa} = (1-\lambda_f)\ IIDM(\{\widetilde{\boldsymbol{v}}_l\}_{l=1}^{K}) + \lambda_f\ IIDM(\{\boldsymbol{f}_l\}_{l=1}^{K}) \tag{5}$$

where λ_f is hyper-parameter to balance these two parts. By aligning the fused features, we can indirectly match the textual feature distributions to minimize the loss of semantic information, while the redundant visual feature can be aligned twice before and after considering the textual information to achieve sufficient alignment.

Domain Augmentation by Dirichlet-Mixup. To compensates for the missing information and increases the data diversity, we adopt Dirichlet-Mixup (Shu et al., 2021) to mix multiple source domains for generating inter-domain data. This method sample the weights from the Dirichlet distribution rather than the beta distribution used in the original mixup (Zhang et al., 2017). Given latent representations of samples from different domain $\{\boldsymbol{h}_i\}_{i=1}^{K}$ where $\boldsymbol{h}_i \in \mathbb{R}^{n \times d}$, where n denotes the number of sample and d is the feature dimension. The mixed feature representation $\boldsymbol{h}_m$ can be calculated as:

$$\delta = Dirichlet(\beta_1, \ldots, \beta_K) \tag{6}$$

$$\boldsymbol{h}_m = \sum_{i=1}^{K} \delta^{(i)} \boldsymbol{h}_i \tag{7}$$

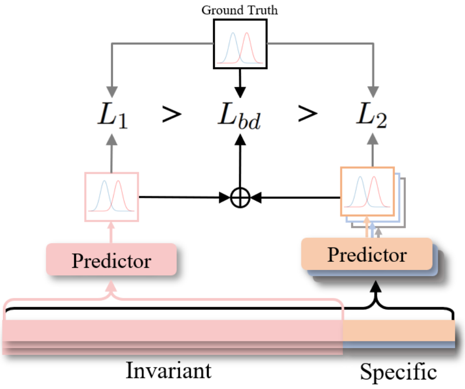


Figure 3: Illustration of our decomposition strategy.

where $\delta \in \mathbb{R}^{K\times n}$. Considering the similarity between the target domain and each source domain is unpredictable, we set the weight β_i of each domain to be equal, which indicates the sampled values $\delta^{(i)}$ of each source domain are statistically equal.

3.3 Features Refinement by Decomposition

Our multi-modal domain alignment method is designed to facilitate the acquisition of domain-invariant representations by the model. However, the NLVL task is characterized by complex structures and multi-step reasoning, with some samples relying heavily on domain-specific information for reasoning. In such cases, learning domain-invariant representations is at odds with learning NLVL reasoning, leading to unstable domain-invariant representations and an unreliable boundary predictor.

To mitigate this conflict between task loss and alignment loss, we introduce the domain-specific boundary predictors $pb^{l^K} l = 1$ for each source domain, as well as a domain-agnostic boundary predictor p_b for all domains. This allows for the dynamic adjustment of weights for these samples and the refinement of features utilized by the domain-agnostic prediction through decomposition.

During training, the boundary predictions for calculating NLVL task loss is the combination of the two kinds of predictors:

$$\{\boldsymbol{T}_s, \boldsymbol{T}_e\} = (1-\gamma)p_b(\boldsymbol{f}_l) + \gamma p_b^l(\boldsymbol{f}_l) \qquad (8)$$

Intuitively, the domain-agnostic predictor that fits all domains tends to use invariant features, while domain-specific predictors prefer to use both invariant and specific information for more accurate predictions. Therefore, as shown in Figure 3, the former's loss with the ground truth L_1 is higher than the latter's L_2, and loss L_{bd} of the combination is naturally between L_1 and L_2.

The key to Eq.8 is to dynamically alter the weight of each sample according to the similarity of these two kinds of prediction. For the hard-to-transfer samples, which mainly rely on domain-specific information for reasoning, their L_1 would be much higher than L_2, and the L_{bd} would also be typically lower than L_1. Accordingly, since gradients are generally proportional to losses, the importance of hard-to-transfer samples and the instability brought by them will be suppressed. On the contrary, for samples, the more their reasoning depends on invariant features, the closer L_1, L_{bd} and L_2 are, and the less their weights are suppressed.

By reducing the importance of the hard-to-transfer samples, the domain-agnostic predictor can learn more stable predictions from invariant features, while no need to fit on specific features. Thus, for the features used for domain-agnostic prediction, the remaining domain-specific components in the learned representations are further decomposed by the domain-specific predictors.

3.4 Training and Inference

For obtaining the scene-robust NLVL model, the final loss function should contain two components: NLVL task loss L_{task} and multi-modal alignment loss L_{maa}. For the NLVL task loss, we apply a similar loss function as used in (Zhang et al., 2020a), which consists of the boundary loss L_{bd} for training a reliable boundary predictor, and confidence coefficient loss L_{cc} to assist in learning discriminative features. The NLVL task loss is a linear combination of the above two sub-losses, *i.e.*, $L_{task} = L_{cc} + 0.2L_{bd}$. The weight is empirically set for balancing the two terms.

Finally, the overall loss function in the training process can be summarized as

$$L = L_{task} + \lambda_{maa} L_{maa} \qquad (9)$$

where the λ_{maa} is the hyper-parameter that depends on the distribution gap in different domains to balance alignment and task loss.

During inference, we only use the domain-agnostic boundary predictor, and the predicted timestamp $\{\tau_s, \tau_e\}$ is determined by the boundary prediction $\{\boldsymbol{T}_s, \boldsymbol{T}_e\}$, which can be written as

$$\{\tau_s, \tau_e\} = \underset{(\tau_s, \tau_e)}{\arg\max}(\boldsymbol{T}_s(\tau_s), \boldsymbol{T}_e(\tau_e)) \qquad (10)$$

Table 1: Performance on the Charades-STA for the scene-robust NLVL task.

Method	Liv.	Bath.	Kit.	Bed.	Avg
Base	33.92	29.31	30.86	30.71	31.20
MMD	32.03	35.19	40.11	33.20	35.13
DANN	35.16	36.89	37.56	34.67	36.07
JSD	33.20	37.63	39.53	32.64	35.75
MMD-AAE	33.01	39.15	38.19	32.00	35.59
Supervised	36.39	44.52	39.14	37.95	39.50
Ours	**35.29**	**39.64**	**40.19**	**35.21**	**37.58**

Table 2: Performance on the ActivityNet-Caption for the scene-robust NLVL problem.

Method	E/D	Pc	Ho	So	Sp	Avg
Base	32.29	35.24	34.51	37.97	33.00	34.60
MMD	32.99	35.25	35.08	39.63	33.06	35.20
DANN	32.44	35.71	34.62	38.67	32.88	34.86
JSD	33.39	37.03	34.25	39.50	32.23	35.28
MMD-AAE	33.31	35.91	35.54	39.37	33.56	35.54
Supervised	36.24	36.74	39.09	42.55	39.12	38.75
Ours	**34.46**	**38.46**	**38.62**	**44.29**	**38.13**	**38.79**

Table 3: Performance on the YouCook2 for the scene-robust NLVL problem.

Method	Am	EA	SA	Eu	Avg
Base	14.13	13.26	9.88	11.85	12.28
MMD	14.82	13.52	10.38	11.71	12.61
DANN	14.55	13.21	10.50	11.67	12.48
JSD	13.88	**13.98**	9.60	**12.21**	12.42
MMD-AAE	14.28	13.31	10.37	12.08	12.51
Supervised	15.85	16.42	11.81	14.11	14.55
Ours	**15.48**	13.72	**11.41**	11.92	**13.13**

4 Experiment

4.1 Dataset

We reconstruct three NLVL public datasets for the scene-robust NLVL task based on their original video tags and perform experiments to evaluate the effectiveness of our framework.

Charades-STA. This dataset is generated by (Gao et al., 2017) from the original Charades dataset (Sigurdsson et al., 2016), which is mainly about indoor activities.

ActivityNet-Caption. It is constructed by (Krishna et al., 2017) and contains about 20,000 untrimmed videos of open activities from ActivityNet (Caba Heilbron et al., 2015).

YouCook2. This dataset (Zhou et al., 2018) includes 2000 long untrimmed videos about cooking. It shows about 89 recipes in 14K video clips. Each video clip is annotated with one sentence.

Based on the video tags provided by the original dataset annotators, each NLVL datasets have different-grained scene splits. For Charades-STA, it can be split by activity location, i.e., *Living room(Liv)*, *Bathroom(Bath)*, *Kitchen(Kit)* and *Bedroom(Bed)*. ActivityNet-Caption is divided by activity event, i.e., *Eat/Drink(E/D)*, *Personal care(Pc)*, *Household(Ho)*, *Social(So)* and *Sport(Sp)*. As for YouCook2, it is split according to the origin of the used recipes, i.e., *America(Am)*, *East Asia(EA)*, *South Asia(SA)* and *European(Eu)*.

With these three restructured NLVL datasets, we iteratively use each domain as the target domain and the remaining domains as the source domains to construct the scene-robust NLVL task. The train/val/test split follows previous works (Zhang et al., 2020a; Zhou et al., 2018) and the scene-robust performance is evaluated on the test set of the target domain.

4.2 Experimental Settings

Implementation Details. We utilize the AdamW (Loshchilov and Hutter, 2017) optimizer and CosineAnnealing scheduler (Loshchilov and Hutter, 2016) with weight decay $1e-6$, and learning rate $5e-4$ for ActivityNet-Caption and $2e-4$ for Charades-STA and YouCook2. During training, the λ_f in Eq.5 is set to 0.2, the γ in Eq.8 is set to 0.1, and the λ_{maa} in Eq.9 is set to 4, 1, 0.1 for ActivityNet-Caption, Charades-STA and YouCook2 respectively. Due to the simple architecture of VSLBase (Zhang et al., 2020a), it can be viewed as a standard proposal-free model. Therefore, our basic network structure is the same as VSLBase to minimize the impact of architecture bias. Please refer to the appendix for more details.

Evaluation Metrics. We adopt "$R@n, IoU = m$" as the evaluation metrics, following (Zhang et al., 2020a). This metric denotes the percentage of language queries having at least one result whose Intersection over Union(IoU) with ground truth is larger than m in the top-n grounding results. In our experiments, we use $n = 1$ and $m = 0.5$.

Baseline. For a comprehensive comparison, we consider the following methods as baselines: 1) Variants of our model, including **Base**, which do not use any alignment and decomposition methods during training, and **Supervised**, which is

Table 4: Framework components ablation on ActivityNet-Caption.

$IIDM_F$	$IIDM_V$	D-Mixup	FRD	Avg
				34.60
✓				36.22
✓	✓			38.26
✓	✓	✓		38.61
✓	✓	✓	✓	**38.79**

Table 5: Intra- and Inter-sample distance metric ablation on ActivityNet-Caption.

Method	Target domain					Avg
	E/D	Pc	Ho	So	Sp	
w/o Intra	34.36	37.76	36.24	43.64	37.44	37.89
w/o Inter	34.11	36.71	35.99	41.17	33.97	36.39
Ours	**34.46**	**38.46**	**38.62**	**44.29**	**38.13**	**38.79**

Table 6: Multi-modal alignment losses ablation on ActivityNet-Caption.

Method	Target domain					Avg
	E/D	Pc	Ho	So	Sp	
$None$	32.29	35.24	34.51	37.97	33.01	34.60
$IIDM_T$	30.91	35.24	33.43	37.12	31.60	33.66
$IIDM_F$	33.77	36.71	35.91	40.98	33.75	36.22
$IIDM_V$	**35.05**	**37.59**	**37.03**	**43.53**	**37.66**	**38.17**

trained on all domains without alignment and decomposition. 2) domain alignment methods for single-modal tasks, including **DANN** (Matsuura and Harada, 2020) and **MMD** (Li et al., 2018b). 3) domain alignment methods for other multi-modal tasks, including **JSD** (Chao et al., 2018) and **MMA** (Xu et al., 2019b).

4.3 Performance Comparison

The quantitative evaluation results of our proposed method on Charades-STA, ActivityNet-Caption and YouCook2 are reported in Table 1, 2 and 3, respectively. The best results are in **bold** and second best underlined. According to the results, we have the following observations:

- On all the benchmark datasets, our method gains noticeable performance improvements compared to the base model, which demonstrates that the proposed methods can effectively help the NLVL model learn the generalizable domain-invariant representations. Besides, stable improvement under different-grained scene shifts is a significant and practical merit since the scene shift with source domains in real-world applications is diverse and unpredictable.

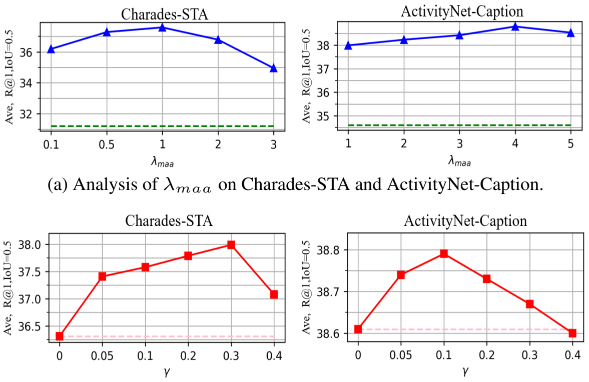
- Remarkably, our method boosts the performance (Avg) of the Base model from 31.20/34.60 to 38.62/38.79 on Charades-STA and ActivityNet-Caption datasets, which far exceeds all the compared methods and even achieves comparable performance to the supervised setting. It further reveals the superiority of our methods on the scene-robust NLVL problem.

- Looking at the evaluation results on YouCook2, the improvement brought by our methods is relatively less than the other two datasets, which may stem from the intrinsic characteristics of this dataset. NLVL on YouCook2 is more challenging than Charades-STA and ActivityNet-Caption. The annotations of YouCook2 are more detailed, and the differences between adjacent frames are slight. The more complex NLVL reasoning on YouCook2 makes it harder to capture discriminative and domain-invariant representations.

4.4 Ablation Study

Component Ablation. We ablate the major components of our framework on Activity-Net Caption: fused features alignment($IIDM_F$), visual features alignment($IIDM_V$), Dirichlet-Mixup(D-Mixup) and feature refining by decomposition(FRD). Results are reported in Table 4. By adding each of our proposed components in turns, the average accuracy gradually increases from 34.60 to 38.79, and each component brings noticeable improvement. The increasing accuracy indicates the effectiveness of each proposed component. In addition, as shown in Figure 4, FRD is even more effective on Charades-STA, with performance improving from 36.31 to 37.58 by default γ, and accuracy can further improve to nearly 38 by adjusting γ.

Design of Distance Metric. The distance metric is critical for learning high-quality domain-invariant representations. To verify the complementarity of the intra-sample and inter-sample distribution in our proposed distance metric, we individually remove the two related term in Eq. 3. As shown in Table 5, the joint use of intra-sample and



(a) Analysis of λ_{maa} on Charades-STA and ActivityNet-Caption.

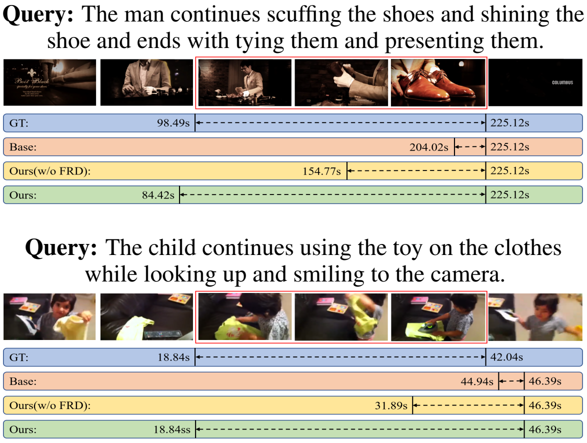
(b) Analysis of γ on Charades-STA and ActivityNet-Caption.

Figure 4: Effect of important hyper-parameters λ_{maa} and γ on Charades-STA and ActivityNet-Caption.

inter-sample distributions leads to the best performance. In addition, the inter-sample part is more critical than the intra-sample part since multiple samples can better reflect the overall distribution of the domain.

Design of Multi-modal Alignment. As discussed in Sec 3.2, the different information density of each modality requires a modality-specific alignment strategy. We separately align features of each modality to prove the necessity and effectiveness of our asymmetric multi-modal alignment strategy. From Table 6, we conclude: 1) Directly aligning information-dense textual features results in performance degradation and loss of semantics. 2) Only aligning the heavily redundant visual feature achieves the best results. 3) Aligning fused features, which can be viewed as symmetrical indirect multi-modal alignment, leads to sub-optimal performance due to insufficient visual information alignment.

Hyper-Parameter Analysis. In our framework, the hyper-parameter γ in Eq. 8 and λ_{maa} in Eq. 9 are important for generalization. Therefore, we further explore their impacts on Charades-STA and ActivityNet-Caption. As shown in Figure 4, for γ, a value of around 0.1 can obtain considerable gain, and too large γ would affect the training of the domain-agnostic predictor. As for λ_{maa}, which depends on the ratio of the alignment loss, the optimal value relies on the distribution gaps between source domains. The bigger domain gaps, the larger λ_{maa} should be set to reinforce alignments adaptively, and vice versa.



Query: The man continues scuffing the shoes and shining the shoe and ends with tying them and presenting them.
GT:
98.49s
225.12s
Base:
204.02s
225.12s
Ours(w/o FRD):
154.77s
225.12s
Ours:
84.42s
225.12s
Query: The child continues using the toy on the clothes while looking up and smiling to the camera.
GT:
18.84s
42.04s
Base:
44.94s
46.39s
Ours(w/o FRD):
31.89s
46.39s
Ours:
18.84ss
46.39s

Figure 5: Qualitative results sampled from the *household* scene in ActivityNet-Caption.

4.5 Qualitative Analysis

In order to qualitatively evaluate the performance of our alignment and decomposition strategy, we show two representative examples from the *household* scene in ActivityNet-Caption, which can be found in Figure 5. (The analysis of failure cases can be found in the appendix.) In both cases, the base model only learns to localize some simple and general actions in the novel scene, such as "smiling at the camera" and "presenting them." By introducing our multi-modal domain alignment method, the representations are forced to be domain-invariant, and the model learns to capture the high-level semantic similarities in different scenes instead of the common overlapping actions. Further, the decomposition approach refines the domain-invariant representations and stabilizes the learning process. Our scene-robust NLVL method can effectively improve localization accuracy on unseen scenes.

5 Conclusion

In this paper, we first proposed a scene-robust problem in NLVL. Our main idea is to learn a group of high-quality domain-invariant feature representations from multiple source domains. By analyzing the extra generalization challenges posed by the NLVL task, we propose a novel NLVL framework that tackles the scene-robust problem from aligning and decoupling perspectives. With the help of these two branches, we significantly enhance the generalization ability to new scenes. Extensive experiments and detailed ablation studies on three widely-used benchmark datasets demonstrate the effectiveness and robustness of our methods.

6 Limitations

In this work, we first formulate the scene-robust NLVL problem and propose our solution. However, our generalizable NLVL model is still tested on existing close-world datasets, and the actual performance in real-world scenarios needs to be further explored. A real-world, large-scale dataset is required to develop a practical, generalized, open-world query-based video retrieval model.

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.

References

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 2017. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232. PMLR.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

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.

Junbum Cha, Hancheol Cho, Kyungjae Lee, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Domain generalization needs stochastic weight averaging for robustness on domain shifts. *arXiv e-prints*, pages arXiv–2102.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5725.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182.

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*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430.

Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018b. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409.

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

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. 2020. Towards recognizing unseen categories in unseen domains. In *European Conference on Computer Vision*, pages 466–483. Springer.

Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. 2018. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1353–1357. IEEE.

Toshihiko Matsuura and Tatsuya Harada. 2020. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11749–11756.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819.

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.

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. 2020. Correlation-aware adversarial domain adaptation and generalization. *Pattern Recognition*, 100:107124.

Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2464–2473.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. 2021. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. 2020. Frustratingly simple domain generalization via image stylization. *arXiv preprint arXiv:2006.11207*.

Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12168–12175.

Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343.

Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019a. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069.

Huijuan Xu, Kun He, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2018. Text-to-clip video retrieval with early fusion and re-captioning. *arXiv preprint arXiv:1804.05113*, 2(6):7.

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2019b. Open-ended visual question answering by multi-modal domain adaptation. *arXiv preprint arXiv:1911.04058*.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*, volume 2.

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019a. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019b. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. 2021. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4206.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

A Analysis on Failure Cases

To better understand the limitation of our framework, we elaborate on the failure case and discuss them in detail. The reasons for our prediction errors can be roughly summarized into three categories:

- **Ambiguity of ground truth.** Due to the complexity of video content, some queries in NLVL datasets correspond to multiple video clips, but the ground truth only label one of them. As shown in Case(1,2) in Figure 6, our predicted segment and the ground truth annotation are both semantically matched to the given query sentence. However, our predictions are considered to be completely wrong.
- **Inability to distinguish subtle actions in video.** In some scenarios, tiny imperceptible action differences in the video is critical to distinguish similar clips. As shown in Case (3), in our predicted segment, the man is actually *refining the draw by paper and a spatula*, rather than *painting*. It might be essential to use a better vision encoder or hierarchical visual feature maps to distinguish the video's subtle actions.
- **Inability to understand detail word.** In the scene-robust NLVL problem, there are inevitable distribution gaps in the vocabulary lists of descriptions in different scenes, which results in misunderstanding some keywords. In Case (4), our model fails to figure out the detailed semantic differences between *solution* and *water* without any information or knowledge about the target domain vocabulary.

Dataset	Scene	#Videos	#Annotations
Charades-STA	Living room	974	2,355
	Bathroom	466	1,157
	Kitchen	1,091	2,663
	Bedroom	1,050	2,581
ActivityNet-caption	Eat/Drink	967	3,962
	Personal care	1,212	4,665
	Household	2,939	10,992
	Social	2,918	10,155
	Sport	6,890	25,152
YouCook2	America	3,800	27,373
	East Asia	2,967	22,561
	South Asia	2,235	19,667
	European	4,274	34,199

Table 7: Statistics of the three NLVL datasets for scene-robust problem.

B Distribution Gaps between Scenes

In this section, we explore the differences in text distribution scenes. Since the videos and sentences in NLVL are semantically matched, the word distribution gaps can also be viewed as semantic gaps between videos in different scenes.

We analyze the word distribution gaps from two perspectives: **Word ratio** and **Vocabulary IoU**.

Word ratio. For each scene, we counted the ratio of occurrences of each word in all sentences. In the upper part of Figure 9, 10, 11, we visualize the word ratios for the same words in different scenes of Charades-STA, ActivityNet-Caption, and YouCook2. The further a word is from the diagonal, the more scene-specific it is. Except for several words with high ratios that are common to all scenes, most words are various widely in word ratio across scenes, as shown by the points off the diagonal.

Vocabulary IoU. In the lower part of Figure 9, 10, 11, we show the Intersection over Union(IoU) of the words with the $topk\%$ words ratio. For sentences, common words(such as: "the," "and," and "is," et al.) appear most frequently. Therefore when $k\%$ is very small, the vocabulary overlap between the two domains is relatively high. However, the IoU decreases sharply as the $k\%$ value slightly increases, which indicates that each scene may contain unique high-frequency words. Note that for Charades-STA, ActivityNet-Caption, and YouCook2, the average IoUs for the entire vocabulary in different scenes are only 0.29, 0.45, and 0.40, respectively. The low IoU value across different scenes on all three figures illustrates that the distribution gap brought by scene shift is prevalent and significant. This observation further demonstrates the necessity and practicality of our scene-robust NLVL problem formulation.

C Implementation Details

Data Processing. For language query, we use the pre-trained Glove (Pennington et al., 2014) embedding to initialize each lowercase word, and the visual embeddings are extracted via the 3D ConvNet pre-trained on Kinetics dataset (Carreira and Zisserman, 2017) as previous method (Zhang et al., 2020a). Note that all the pre-trained feature extractors are fixed during training. All experiments were carried out on a single 2080ti.

Network Architecture. Due to the simplicity and effectiveness of VSLBase (Zhang et al., 2020a), it can be viewed as a standard proposal-free NLVL model. Therefore, our network structure is similar to VSLNet to minimize the impact of architectural bias. The video and query feature encoder consist of four convolution layers and a multi-head attention layer. After feature encoding, we use context-query attention(CQA) as our cross-modal representation fusion module. The boundary and confidence coefficient predictors are essentially multi-layer perceptrons (MLP). We set the kernel size of the convolution layer as 7 and the head size of multi-head attention as 8. The frame number of video N_v is set to 32 for all three datasets. And the hidden dimensions are set to 128 for Charades-STA, and 256 for YouCook2 and ActivityNet-Caption.

D Ethical Discussion

Natural language video localization has many applications, including video corpus retrieval and video-

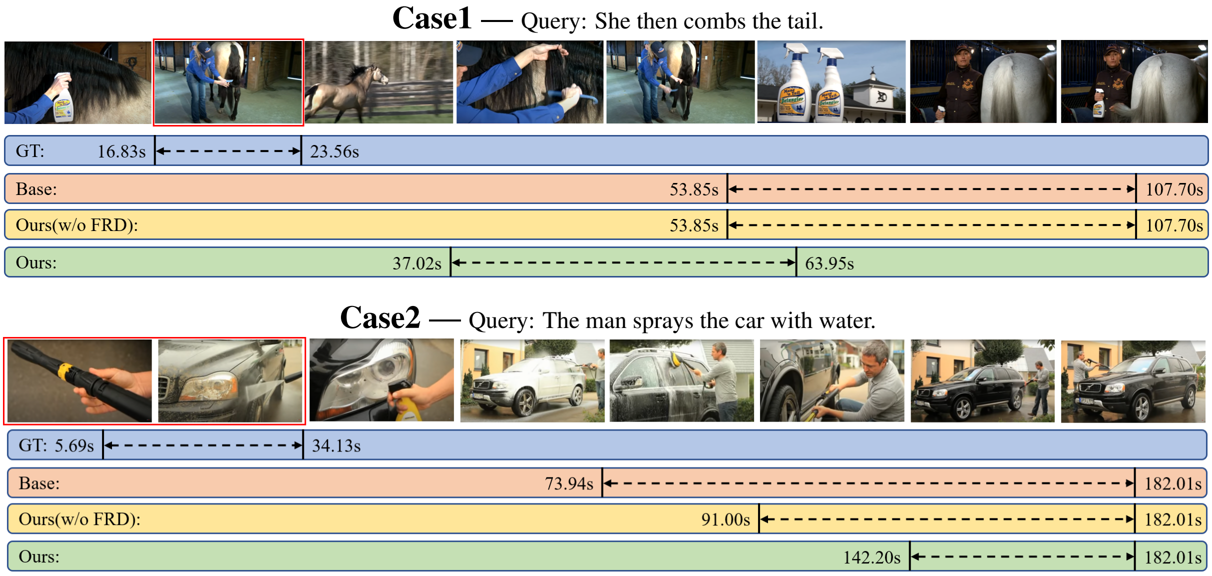

Case1 — Query: She then combs the tail.
GT: 16.83s 23.56s
Base: 53.85s 107.70s
Ours(w/o FRD): 53.85s 107.70s
Ours: 37.02s 63.95s
Case2 — Query: The man sprays the car with water.
GT: 5.69s 34.13s
Base: 73.94s 182.01s
Ours(w/o FRD): 91.00s 182.01s
Ours: 142.20s 182.01s

Figure 6: Failure Cases – Ambiguity of ground truth.

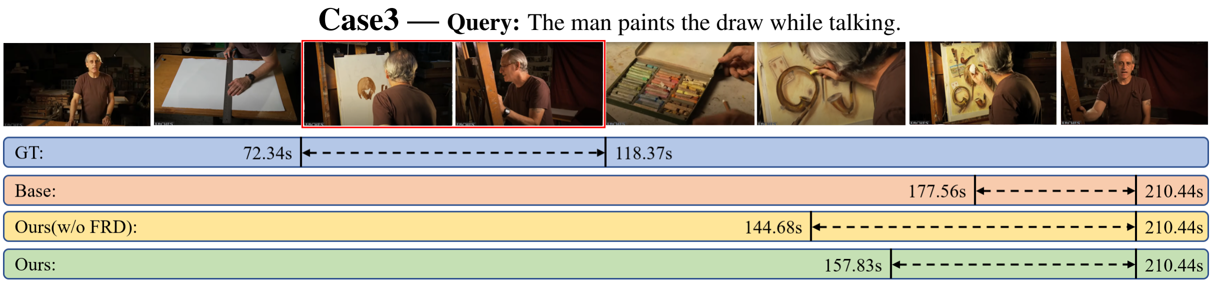


Figure 7: Failure Cases – Inability to distinguish subtle actions in video.

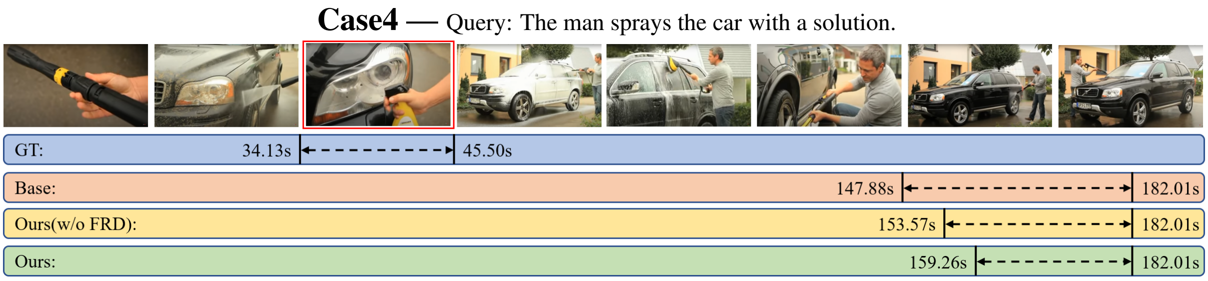


Figure 8: Failure Cases – Inability to understand detail word

based dialogue. Our scene-robust NLVL problem makes it more practical and reliable in real-world prediction scenarios. Due to the generalization of our method, there may be concerns about the misuse of offensive data. However, in fact, our method focuses on the scene shift while the general activities are consistent, such as Charades-STA for indoor activity, ActivityNet-Caption for outdoor activity, and YouCook2 for cooking. Therefore, the model may not get reliable generalization performance on entirely new activities.

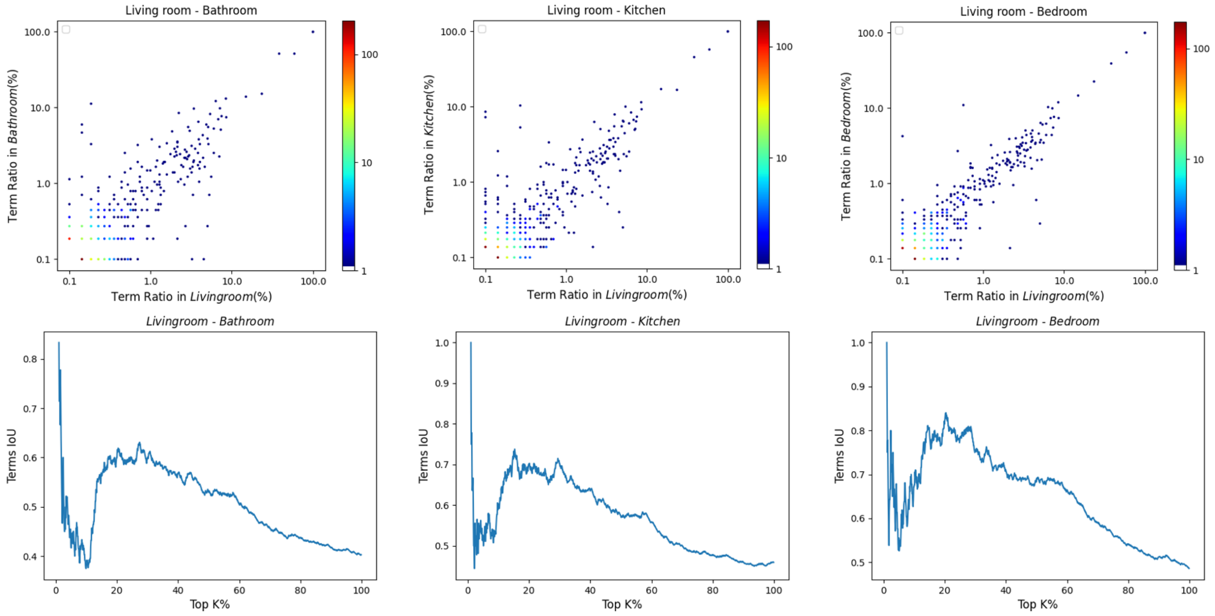


Figure 9: Word distribution statistics on Charades-STA.

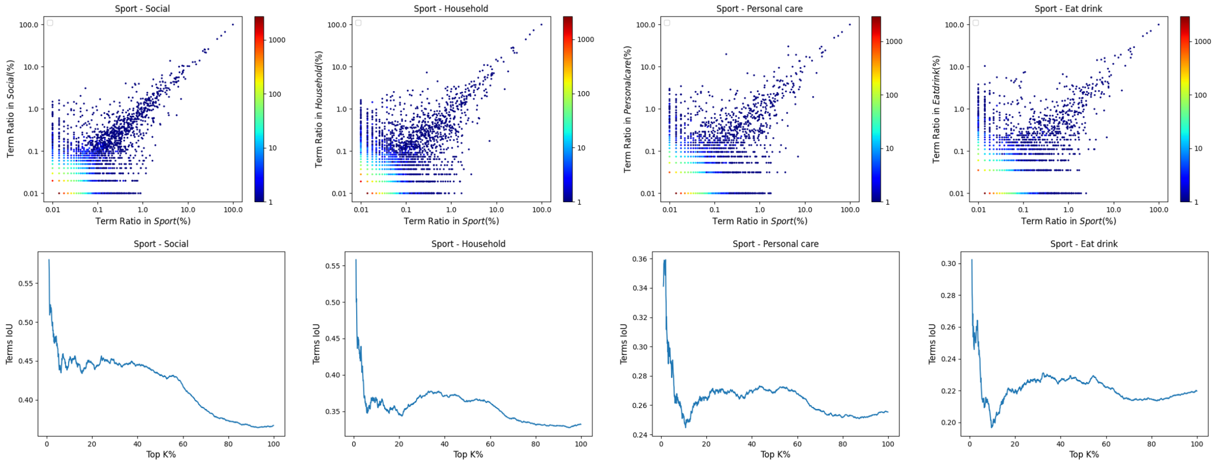


Figure 10: Word distribution statistics on ActivityNet-Caption.

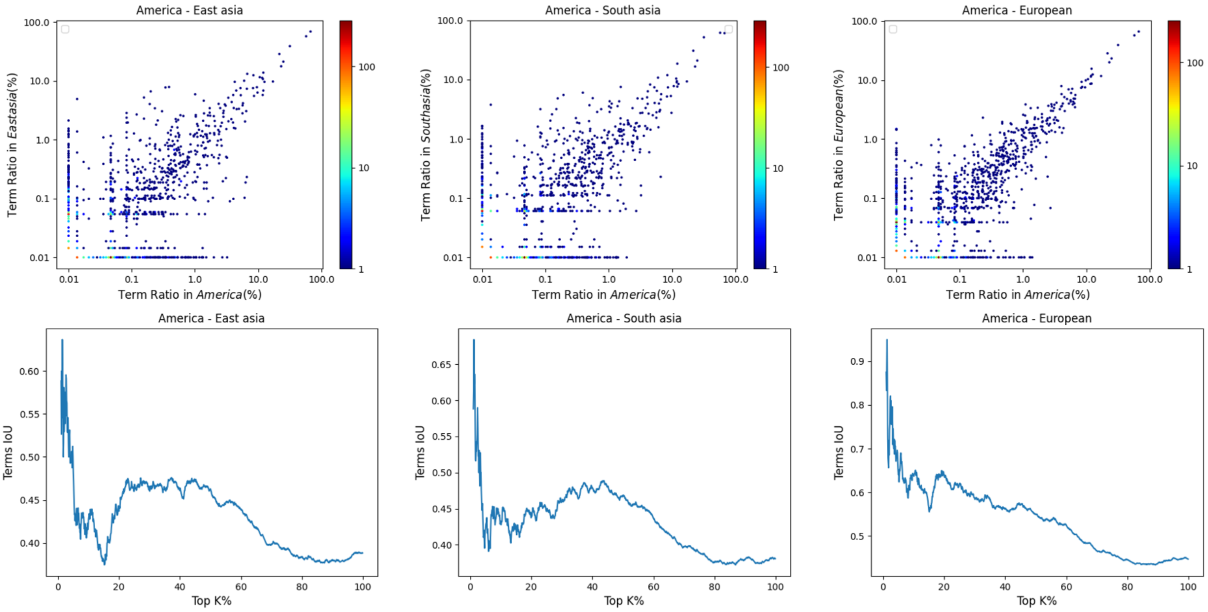


Figure 11: Word distribution statistics on YouCook2.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the Sec 6 and Appendix.A

☑ A2. Did you discuss any potential risks of your work?
In the Appendix.D

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

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

B ☑ Did you use or create scientific artifacts?

In the 4.2 and the Appendix.C

☑ B1. Did you cite the creators of artifacts you used?
In the reference

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In the reference

□ 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)?
Not applicable. Left blank.

□ 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?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ 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.
In the 4.1 Dataset

C ☑ Did you run computational experiments?

In the sec 4 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In the sec 4.2 and Appendix.c

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?
In the sec 4.2, 4.4 and Appendix.c

□ 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?
Not applicable. Left blank.

☑ 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.)?
In the sec 4.2 and Appendix.c

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.?
Not applicable. Left blank.

□ 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)?
Not applicable. Left blank.

□ 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?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.