A Multi-level Alignment Training Scheme for Video-and-Language Grounding

Yubo Zhang
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
zhangyb@cs.unc.edu

Feiyang Niu
Amazon Alexa AI
Sunnyvale, CA, USA
nfeiyan@amazon.com

Qing Ping
Amazon Alexa AI
Sunnyvale, CA, USA
pingqing@amazon.com

Govind Thattai
Amazon Alexa AI
Sunnyvale, CA, USA
thattg@amazon.com

***Abstract*—To solve video-and-language grounding tasks, the key is for the network to understand the connection between the two modalities. For a pair of video and language description, their semantic relation is reflected by their encodings' similarity. A good multi-modality encoder should be able to well capture both inputs' semantics and encode them in the shared feature space where embedding distance gets properly translated into their semantic similarity. In this work, we focused on this semantic connection between video and language, and developed a multi-level alignment training scheme to directly shape the encoding process. Global and segment levels of video-language alignment pairs were designed, based on the information similarity ranging from high-level context to fine-grained semantics. The contrastive loss was used to contrast the encodings' similarities between the positive and negative alignment pairs, and to ensure the network is trained in such a way that similar information is encoded closely in the shared feature space while information of different semantics is kept apart. Our multi-level alignment training can be applied to various video-and-language grounding tasks. Together with the task-specific training loss, our framework achieved comparable performance as previous state-of-the-arts on multiple video QA and retrieval datasets.**

***Index Terms*—video-and-language grounding, multi-level alignment, contrastive learning**

I. Introduction

Having a machine that can follow human instructions and adapt to visual surroundings is the key to building a intelligent system to aid human beings in our daily activities. To achieve this goal, the smart system will need to understand the meaning of the input natural language, and to be aware of the information embedded in the visual input such as videos. More importantly, the system has to have the ability to make connections between the two modalities to further reason with the joint-modality information.

Neural network models have been proved to be powerful for understanding complex real-world information, and have shown their strength in solving video-and-language grounding problems, such as video question answering (QA) and video retrieval. Examples of three common video-and-language grounding tasks are included in Figure 2. Task-specific frameworks have been developed, targeting each grounding task's own challenges. Some auxiliary techniques, e.g., object detection and scene graph reasoning, have been explored in order to achieve better performance [1]–[5]. However, these tailored models are usually lacking generalizability, and serious modifications are often required for them to handle other types of grounding tasks [6], [7].

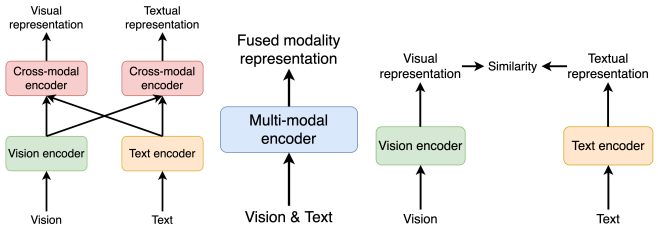
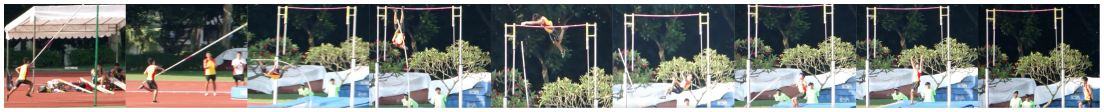


Fig. 1: Three types of visual-linguistic model architectures for cross-modal learning: (a) Cross-modality type, (b) Joint-modality type, and (c) Separate-modality type

Recently, following the success in natural language processing (NLP) tasks [8], large-scale BERT-type models have been widely applied to vision-and-language grounding problems [4], [9]–[13]. The models are usually pre-trained with a large amount of data in both modalities, with standalone tasks that emphasize networks' general ability to understand language and visual information and their interconnection, such as masked language/frame modeling. When applied to the specific grounding task, the models are fine-tuned with the task's own data to generate certain output, while their ability to reason upon multi-modality input is reserved. BERT-type models are versatile and powerful, but considering their scale, the training process is usually time-consuming and computationally challenging. Moreover, compared to task-specific frameworks, the reasoning flow between the two modalities is less well represented inside BERT.

By observing the existing visual-linguistic models, we categorize their inter-modality reasoning architectures into the following three main types: 1) **cross-modality** (Fig. 1(a)), such as in ViLBERT [9] and LXMERT [14], where visual and textual representations are encoded by two separate transformers and then a multi-modal transformer exchanges cross-modal information through a novel co-attentional layer; 2) **joint-modality** (Fig. 1(b)), such as VL-BERT [15] and VideoBERT [16], who fuses visual and textual information



Question: What's the person in yellow shirt doing?
Candidates: (1) high jump (2) run (3) pole vault (4) play kickball

(a) Video QA

Candidates:
(1) a man is glaring, and someone with sunglasses appears (2) a cat tries to catch a mouse on a tablet
(3) a man dressed in red is dancing (4) a gorilla comes close to another in the zoo

(b) Video-text retrieval

Query: A women is putting a Teddy bear into a box.

(c) Video moment retrieval

Fig. 2: Video-and-language tasks. (a) Video QA task aims to predict answers to natural language questions given a video as context. (b) Video-text retrieval selects the text description from the candidate pool that best matches the video content (vice versa, text-video retrieval). (c) Video moment retrieval localizes the video segment that aligns with the query text input. Ground-truths are colored in blue.

at the initial stage of a joint transformer model; 3) **separate-modality** (Fig. 1(c)), such as COOT [17] and T2VLAD [18], who encodes visual and textual representations with two separate streams, then joints the visual-textual feature space by calculating the inter-modality similarity. Among these three reasoning types, both cross-modality and joint-modality calculate pairwise attention between $\mathcal{M}$ visual patches and $\mathcal{N}$ textual tokens, requiring $\mathcal{O}(MN)$ time complexity of intra-model information exchange. On the other hand, separate-modality type models only require a time complexity of $\mathcal{O}(M+N)$. With better efficiency, it is favored by recent portable video-language applications, and our model falls into this realm.

In this work, we developed a lightweight grounding framework that is versatile for different kinds of video-and-language grounding tasks. To efficiently train the network, we designed a set of multi-level video-language alignment losses, which are built upon contrastive loss with sophisticated designed positive and negative video-language pairs, to directly shape the feature space. In the alignment losses, the features of relevant video and language information are aligned respectively in high to low semantic levels and across different time spans. Using this training scheme, the input from different modalities representing similar subjects is encoded closely in common the feature space, and this characteristic benefits the downstream video-and-language grounding tasks. When tested on video QA and retrieval tasks, our method achieved state-of-the-art results on multiple datasets, and it also shows a great potential to be applied in other grounding problems such as video moment retrieval.

II. Related Work

A. Vision-and-Language Pre-training

Driven by revolutionary advances in NLP led by transformer-based pre-training frameworks such as BERT [8], GPT2 [19], XLNet [20], and GPT3 [21], recent years have witnessed a boom in the area of extending the use of transformer architectures to the visual-linguistic tasks. Pioneering works such as ViLBERT [9] and LXMERT [14] adopt two separate transformers for image and text encoding independently and propose a novel co-attentional transformer layer to fuse visual and linguistic representations. Inspired by the original BERT pre-training tasks, ViLBERT [9] is trained through the tasks of reconstruction of the masked image region or text tokens and alignment check if the caption describes the image content. VisualBERT [22], Unicoder-VL [23], VL-BERT [15], and UNITER [4] advocate single-stream architecture, where the text and vision sequences are combined as the input of one shared transformer encoder. A new Word-Region Alignment task is proposed by UNITER to further explicitly bridge the fine-grained alignment between visual regions and text tokens. It's worth pointing out that both architectures rely on pre-trained object detectors for extracting ROIs that are viewed as individual visual tokens. A few other works, such as Pixel-BERT [24] and VirTex [25] for images or HERO [11] for

video, operate directly over dense feature maps instead of ROIs extracted by pre-trained object detectors. In these approaches, both visual and textual features are fed into a transformer-based model usually pre-trained with multiple losses. Through these pre-training explorations, tremendous advances have been made in the area of vision-and-language representation learning.

B. Contrastive Learning

Contrastive learning is a framework that learns such an embedding space that similar sample pairs stay close to each other while dissimilar ones are kept far apart. Contrastive learning can be used in both supervised and unsupervised settings. Remarkable progress has been seen in recent studies in unsupervised visual representation learning [26]–[31] leveraging the power of contrastive learning. We review several popular representative contrastive learning methods that benefit from optimization with negative (dissimilar) samples. Typically, samples in the current mini-batch are utilized in a way that its augmented views are considered as positive samples and are paired with other samples in the same batch as negatives. Computing embeddings for a large number of negative samples in every batch could be computationally prohibitive. As a work-around, memory bank [26] was proposed to store representations of all samples in the dataset from past iterations. The dictionary for each mini-batch is randomly sampled from the memory bank with no back-propagation, so it can support a large dictionary size. However, the representations in the memory bank are from very different encoders all over the past epoch and they are less consistent. MoCo [30] provides a framework of unsupervised learning visual representation as a dynamic dictionary look-up. Compared to memory bank, it enables us to reuse representations of immediately preceding mini-batches due to a queue-based dictionary and is more memory-efficient and can be trained on billion-scale data. SimCLR [29] learns visual representations by maximizing agreement between differently augmented views of the same sample via a contrastive loss in a latent space. It advocates large batch size negatives, stronger data augmentation and introduces the learnable nonlinear transformation, altogether helping improve unsupervised visual representation learning. Our method also benefits from the large-scale negative sample learning. The effects of cross-modal learning without negatives are not discussed in this paper.

C. Video-and-Language Tasks

Popular video-and-language tasks include text-video retrieval [32]–[37], video moment retrieval [11], [38]–[40], video captioning [32], [35], [37], [38], [41], video question answering [11], [42]–[46], and video-and-language inference [47]. Text-video retrieval selects a video from a pool of candidate videos, whose content best matches the input text query. Compared to text-image retrieval, text-video retrieval is more challenging that requires the understanding of temporal dynamics and complicated text semantics. Video moment retrieval requires localizing video segments from natural language queries. Video captioning is the task of generating sentences that well describe the input video content, and video question answering aims to predict answers to natural language questions given a video as context. These tasks mainly focus on explicit factual descriptions or explicit information of the video. In contrast, video-and-language inference requires not only explicit visual cues but also more sophisticated reasoning skills, such as inferring reasons and interpreting human emotions. Several efforts [5], [11], [12], [16]–[18], [48], [49] have been made that leverage the powerful transformer architecture as the visual backbones and apply contrastive learning for video-and-language learning tasks. VideoBERT [16] represents a video with a combined sequence of textual tokens and selected video frames and applies a transformer to learn joint representations. ActBERT [5] distinguishes between global actions and local regional objects and encodes them jointly with linguistic descriptions. COOT [17] proposes a hierarchical model that exploits long-range temporal context to produce video-text embedding based on hierarchical interactions between local and global context. Recently, T2VLAD [18] extracts features from the aspects of scene and action, and performs similarity matching with the representations of each local token and the global sentence, while HiT [49] conducts cross-matching between feature-level and semantic-level embedding. But they do not simultaneously decompose the video and text to conduct deep alignment, from where we propose the multi-stream multi-level alignment framework that can be universally applied to various video-and-language tasks.

III. Methods

In this section, we will discuss the proposed multi-level alignment training scheme in detail. The training objectives, which models global-level and segment-level alignments, are designed for the network to capture different levels of semantic connection between language and vision. Together with tasks' specific training loss, the training scheme can be readily applied to various grounding tasks. The overall framework of our method is described in Figure 3.

A. Network Architecture and Notations

Following the lightweight architecture concept, our network adopts the typical encoder-decoder structure and models inter-modality relation using separate-modality design. The language and video inputs are encoded with their own encoders, and then passed on to the task-specific decoder to produce downstream answers.

Specifically, our text encoder is implemented as LSTM layers. It takes tokenized sentence T as the input, and produces every word's encoding E_w. After self-attention, words' weighted sum E_L represents the general semantic of the input paragraph. Meanwhile, the video encoder consists of feature extractors and an extra MLP layers. We apply two streams of feature extractors, specifically, ResNet [50] to extract static object-level visual features and SlowFast [51] to extract temporal motion-level features, and the MLP produces

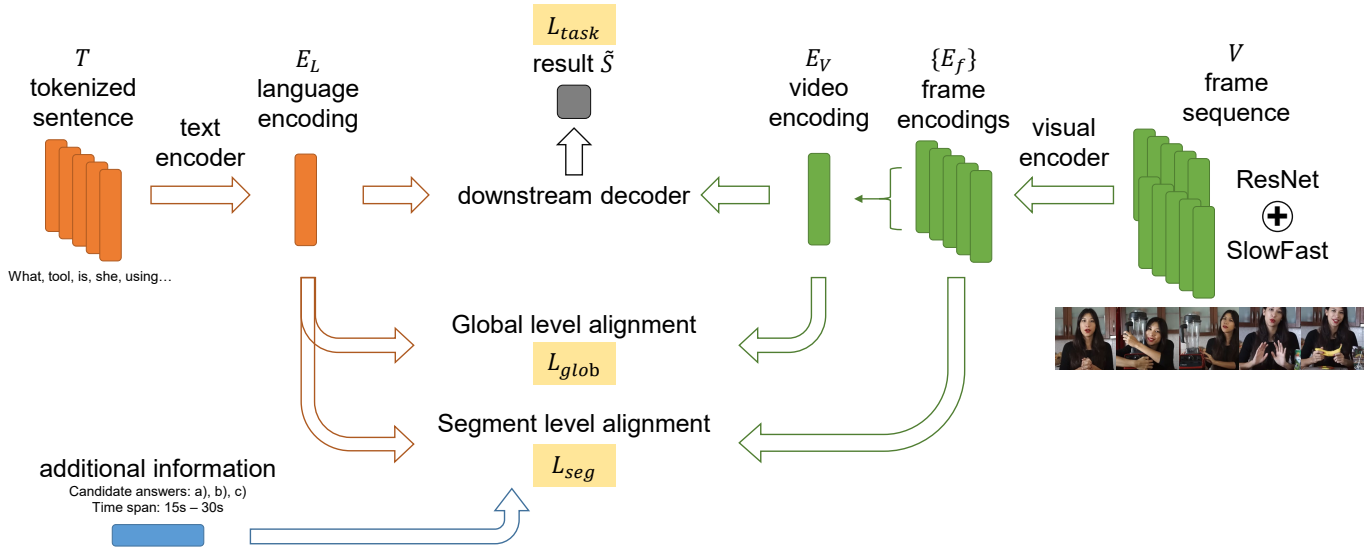


Fig. 3: The lightweight video-and-language grounding network and its multi-level alignment training scheme. The video and language inputs are encoded by their individual encoders, then the downstream decoder produces downstream task results. The network is trained with the task-specific loss and multi-level alignment losses.

each frame's representation E_f from the concatenation of the two. The video encoding E_V is the weighted sum of every frame's E_f, where the weights are computed from a self-attention layer.

For each downstream grounding task, a decoder that generates the task-specific output is designed, which takes the encodings E_L and E_V and produces the output result $\tilde{S}$.

B. Training overview

The input of each downstream grounding task is a pair of language and video samples. These two modalities are semantically related, as their information together infer the result of the task. Thus, in this work we design two levels of semantic alignment losses, L_{glob} and L_{seg}, to enforce the encoders relating the semantic relations between the two modalities, whose details will be discussed in Section III-C.

For every downstream task, the network will also be trained on the task-specific objective L_{task}, e.g., cross-entropy loss for multi-choice video QA task. As the multi-level alignment losses directly shape the encoders, L_{task} is indispensable to train the decoder in order to obtain reasonable results, further tuning the encoders to learn task-specific information.

To sum up, the overall training loss of our framework is the weighted sum of the objectives mentioned above, where λ_i are hyper-parameters to balance individual loss contribution:

$$L_{train} = L_{task} + \lambda_1 \cdot L_{glob} + \lambda_2 \cdot L_{seg} \tag{1}$$

C. Multi-level Alignment

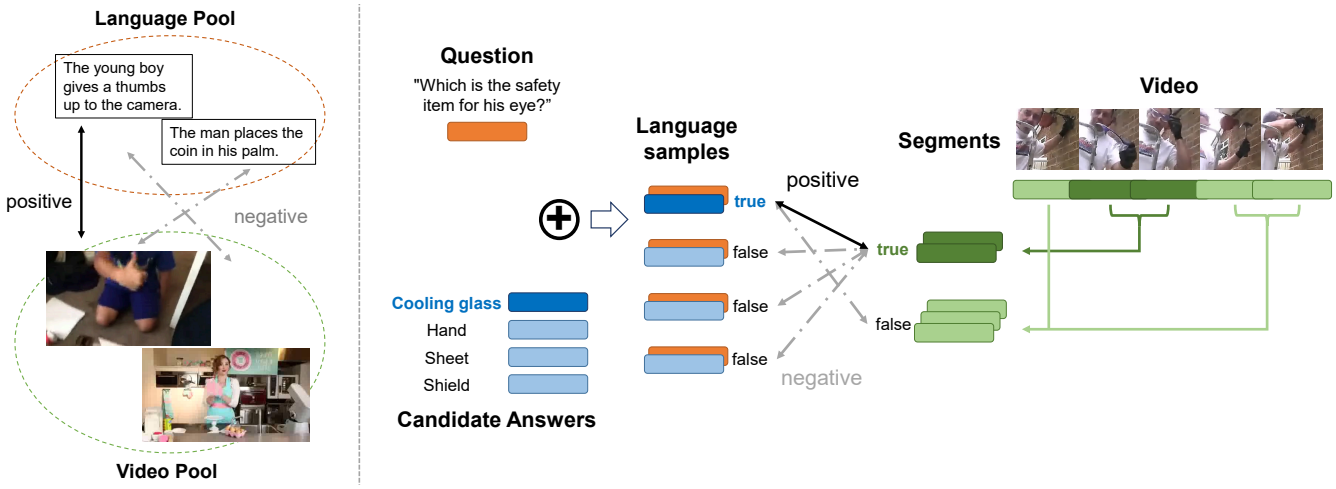
Our alignment losses train the network to identify semantic relations by encoding them properly in the feature space. By contrasting the embedding similarity of modality pairs that are more relevant to the ones that have less semantic connection, the network is able to ground the similar information closer in the shared feature space, bridging the two modalities. Thus, each of our alignment losses is constructed with the contrastive loss function, where α is a pre-defined margin:

$$L(S^{neg}, S^{pos}) = [\alpha + S^{neg} - S^{pos}]_+ \tag{2}$$

It penalizes the similarity score of a negative video-language alignment S^{neg} and encourages a higher similarity score of the positive alignments S^{pos}.

The language and video can be decomposed into word-wise and frame-wise information units. As the units gradually group together, they can convey information from details to overview. Thus, our embedding matching, i.e., the design of positive/negative pairs in contrastive losses, is conducted on multiple semantic levels. The global level alignment loss trains the network to capture high-level global information; while the segment level alignment focuses on fine-grained details of language and video inputs. The positive/negative pairs of these two levels of alignment losses are designed differently; their matching schemes will be discussed in detail below.

1) Global Alignment Loss: The highest level of embedding alignment is conducted globally. L_{glob} is designed for the network to capture the overall semantic relation of a video clip and its language description, potentially benefiting the video-and-text retrieval task. It aligns the corresponding video and language pair from large video and language pools. As shown in Figure 4(a), the language pool contains all the language descriptions of the dataset and the video pool contains all the video clips; meanwhile, each description has a video correspondence as the dataset provided. The embedding similarity of a video-language pair is computed as the cosine distance



(a) Global Alignment

(b) Segment Alignment

Fig. 4: Multi-level alignment: the positive/negative matching scheme of global level and segment level.

$\cos(*,*)$ between the language and video encoding vectors:

$$S_{glob} = \cos(E_L, E_V) \tag{3}$$

The global alignment loss will contrast the similarity score of the positive video-language pair, which both describe the same theme, to the scores of negative pairs that mismatch. Since the choice of positive or negative is relevant to the scale of the entire dataset, in practice we use the batch-wise hardest negative alignment in L_{glob}:

$$L_{glob} = L(\max_{\{i \in batch\}} S_{glob}^{neg_i}, S_{glob}^{pos}) \tag{4}$$

Therefore, with L_{glob}, the overall features of language and video are encoded closely in the feature space if they both represent similarly themed information.

2) Segment Alignment Loss: To train the network to capture more fine-grained alignment, we further introduce the segment level alignment loss L_{seg}. For downstream datasets, ground truth sometimes highlights the details and contains additional information that can be used for further supervision. For instance in video QA tasks, the ground-truth answer can serve as additional knowledge, and the dataset may provide the time-span information on which the question is based. As shown in Figure 4(b), the fine-grained alignment can be constructed using the additional information, benefiting the QA task. The question combining the correct answer and other candidate answers forms the "true" and "false" language samples. Meanwhile, the video clip can also be divided into the "true" segment, which contains the frames that the question is based on, and the "false" segments that are irrelevant. The positive alignments are between the "true" language sample and the frames from the "true" segment, while the negative alignments are between the "false" language sample and "true" frames or vice versa. Here, the similarity score of each alignment is computed as

$$S_{seg} = \cos(\frac{1}{2}(E_L + E_{ans}), E_f) \tag{5}$$

where E_{ans} denotes the answer's encoding.

Compared to the global alignment, the segment level alignment focuses on the interior relation of a single video-language context, so the loss L_{seg} is implemented on the data sample-wise:

$$L_{seg} = L(\max_{\{j \in sample\}} S_{seg}^{neg_j}, S_{seg}^{pos}) \tag{6}$$

The elaborate matching from L_{seg} further tunes the feature space, and benefits the tasks such as video QA and moment retrieval that generally require understanding subtle details.

IV. Experiment

Our framework is versatile for various kinds of video-and-language grounding tasks. We applied our method on two public video QA datasets, How2QA [52] and ActivityNet-QA [45], and a video-text retrieval dataset, TGIF [36], where our multi-level alignment training scheme helps the lightweight network achieve comparable results to previous state-of-the-arts. Besides, we explored the possibility to apply our method to the video moment retrieval task, testing on How2R dataset [52].

A. Implementation Details

For an input video, we used ResNet [50] to extract static visual features frame-by-frame and SlowFast [51] to extract temporal motion features. Both ResNet and SlowFast feature extractors were loaded with pre-trained parameters that were fixed during training. Specifically, we used the implementation of ResNet152 pre-trained on ImageNet data [53] from Torchvision and SlowFast pre-trained on Kinetics [54]. For every

Dataset	Method	Accuracy (%)
How2QA [52]	HERO [11]	60.42
	Ours	**63.11**
ActivityNet-QA [45]	E-SA [45]	31.8
	CoMVT [56]	36.6
	VQA-T [57]	**36.8**
	Ours	36.3

TABLE I: Video QA results: our method compared to the previous state-of-the-arts on the same experimental settings, no subtitle and no pre-training.

Study	Method	Accuracy (%)
Alignments	None	59.84
	Global	60.90
	Segment	**63.11**
	Global + Segment	61.96
Feature extractors	ResNet	62.47
	SlowFast	62.63
	ResNet + SlowFast	**63.11**

TABLE II: Ablation studies: the effect of different alignment loss combinations and feature extractors, tested on How2QA [52].

experiment listed in this section, the network was trained for 50 epochs, with a mini-batch size of 64. We used AdamW [55] to optimize model parameters, with a learning rate $1e-4$, $\beta_1 = 0.9$ and $\beta_2 = 0.98$, weight decay 0.01. Due to the fixed weights of feature extractors, we were able to complete the 50 epochs of either QA or retrieval task training within 1 day on one V100 GPU.

B. Video QA Results

The How2QA dataset [52] contains Youtube instructional videos with their annotated questions, whose corresponding time spans are also given. The answer to each question will be chosen from four candidates. We trained our network on their public training set, which contains 34k questions, and the result in Table I is reported on the public validation set that contains about 3k questions. To get the best result of How2QA dataset, we found our best setting is $\lambda_1 = 0$ and $\lambda_2 = 1.0$ in the alignment scheme Eq. 1, which means only activating the segment matching constraint. Although the dataset also provides the subtitle paragraphs of the videos, we did not include this information in training and testing; when compared to the previous state-of-the-art method HERO [11] in the same setting (without pre-training), our lightweight model achieved better performance than the BERT-type model, with an improvement in accuracy of absolute 2.69%.

The ActivityNet-QA dataset [45] contains 58k QA pairs that come from 5.8k activity videos. The questions are open-ended, but we took the most frequently occurred 1k answers as the candidates. In our segment-level alignment, we randomly picked three "false" answers from the candidates, and as no time span information was provided, no "false" video segment was contrasted. Meanwhile, in the global-level matching, the language pool was the set of question-answer pairs in the batch while the video pool contained the videos. When trained with $\lambda_1 = 1.0$ and $\lambda_2 = 2.0$, our network achieved 36.3% accuracy on the public test set, which is comparable to the performance of previous state-of-the-art methods [56], [57] in the same non-pre-training setting.

1) Ablation Study:

a) Multi-level alignments: To study the effect of our multi-level alignment scheme, on the How2QA dataset we did the ablation study of different loss combinations. As shown in the first section of Table II, when the network is only trained with task-specific loss and no alignment loss was applied, the accuracy is lower than 60%. When one of the alignment losses was added, the performance improved, and the best result came from adding the segment level matching. However, we notice that when both levels of alignment were applied, the performance was slightly worse than applying the segment alignment alone. The possible reason may be the characteristic of this dataset, where the question can only correspond to a small fraction of the video, thus the global alignment may cause confusion with information from irrelevant frames while the network simultaneously learning the more fine-grained segment alignment. This finding informs future applications to activate alignment levels differently according to the feature of their data.

b) Visual features: As we applied two streams of feature extractors in our framework (Fig. 3), we also studied the necessity of using them both. As shown in the second section of Table II, when the network was only given one type of visual features, i.e., ResNet static features or SlowFast motion feature, the performance decreased slightly, indicating the merit of applying both types of features.

2) Alignment Visualization: To further analyze the effect of alignment losses, we visualize the correlation between the encodings of two modalities. In Fig. 5, we show two cases from the How2QA test set, with the results from the network trained without or with alignment losses. The heatmaps show the cosine similarity between the encoding features of each language (question+candidate answer) and frame pair. The brighter the color is, the stronger the vision and language are correlated to each other.

In the first example, the person is showing an orange to the camera at the beginning of the video and near the end. For the network trained without the alignment scheme, it failed to correlate the "holding" action with the corresponding frames of the video, predicting the wrong answer. However, the network trained with our alignments successfully recognized the action in early frames as the similarities there are higher, therefore giving the right answer. Similarly, in the second case, the feature similarity between the correct answer (blue straw) and the frames in the corresponding timespan is the highest when alignment losses were used. On the other hand, without the alignment training, the information from irrelevant frames distracted the network, causing it to predict the straw as white.

The visualization of the feature similarities in these test

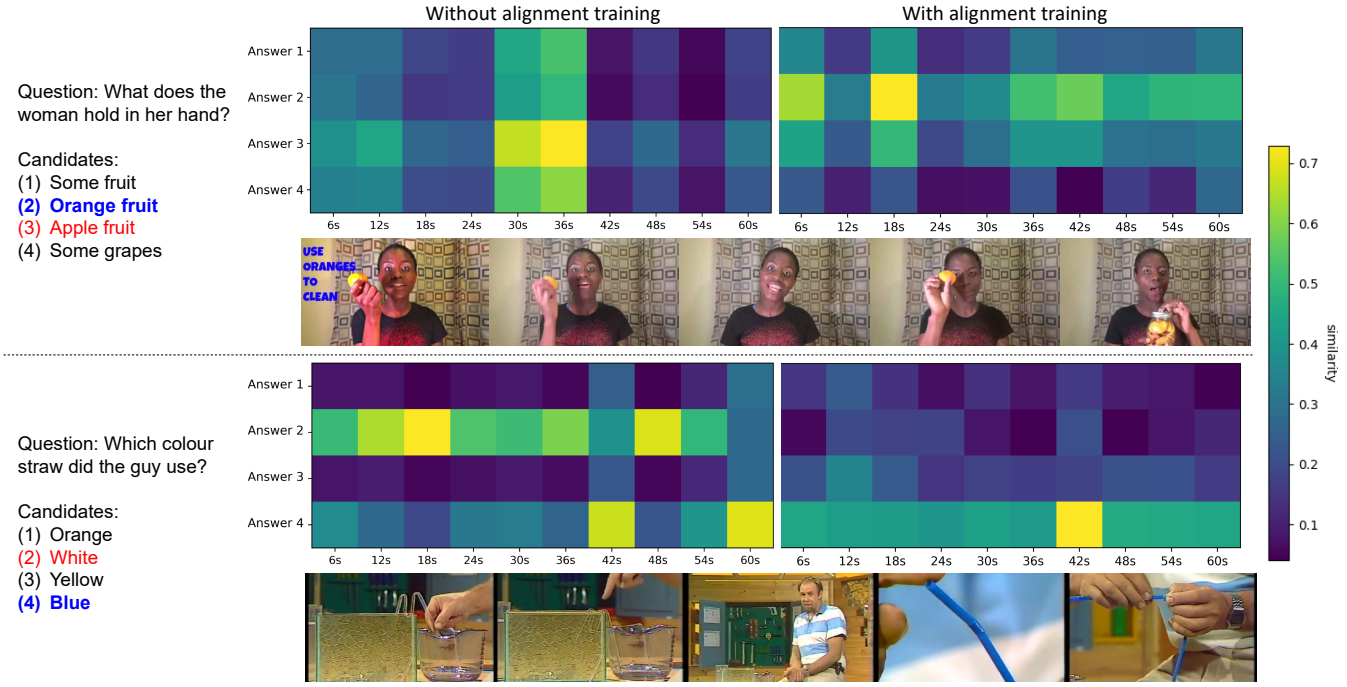


Fig. 5: The heatmaps show the encoding feature vectors' similarity between each pair of question+candidate and video frame. When the network was not supervised with alignment losses, the correlation between two modalities was less clear; wrong answers (marked as red) may get more "attention" at the wrong time, as the heatmaps shows on the left. When the alignment loss was added, the predicted correlation became more reasonable, leading to the correct predictions (marked as blue).

cases clearly shows the benefit of using additional alignment information in the training. The correlation between the two modalities is better learned by the network, leading to better performance on the QA task.

C. Video-text Retrieval Results

Method	Retrieval R@1	
	Text-to-video	Video-to-text
DeViSE [58]	2.2	2.1
Corr-AE [59]	2.1	2.2
Order [60]	1.6	1.7
VSE++ [61]	1.6	1.4
PVSE [62]	3.0	3.3
HGR [6]	4.5	–
Ours	**4.7**	**6.1**

TABLE III: Video-text retrieval results on TGIF dataset [36]. The results of previous works were reported in literatures [6], [62].

Our framework can also be applied to the video-text retrieval task, where the network's job is to find the corresponding text/video of the given video/text. In this application, global level alignment is activated, which alone should be capable for the job; meanwhile, the task-specific loss function is not needed and segment matching is not applicable. We tested our method on the TGIF dataset [36], which contains short gif videos and their descriptions. We used their training set that contains 80k videos for the training, and the results in Table III are reported on the official test set with 11k videos. Compared to the previous methods [6], [62] that were specifically designed for retrieval, our framework can achieve the state-of-the-art result of the recall at the top 1 ranking, showing the capability of our method to be applied to different types of grounding tasks.

D. Video Moment Retrieval Results

Alignment	tIoU ≥ 0.5 (%)	tIoU ≥ 0.7 (%)
None	9.2	2.5
Global	10.9	**3.9**
Segment	10.4	3.0
Global + Segment	**11.0**	3.8

TABLE IV: Our approach's video moment retrieval results on How2R dataset [52], under different alignment settings.

Lastly, we explored the possibility to apply our multi-level alignments on the video moment retrieval task How2R [52], where the description of a certain segment of the video is given, and the network needs to predict the time span of that segment. For our global level alignment, language and video pools were implemented as the collections of the data in the mini-batch; in the segment level, no "false" language sample was contrasted. We used the cross-entropy loss as the task-specific loss for the network to further learn the correct starting and ending timestamps.

Table IV lists our approach's results under different alignment settings, evaluating under temporal Intersection over Union (tIoU) that measures the overlap between the predicted span and the ground-truth span. Compared to the network trained without any alignment loss, the performance improved as we added either global or segment alignment loss to further constrain the training. This indicates the merit of alignment losses in video moment retrieval tasks. However, we notice that using both alignments did not outperform applying only one alignment. This might due to the nature of the dataset, and better hyper-parameter searching to balance global-segment information might potentially improve the performance.

V. Conclusion

In this work, we developed a multi-level alignment training scheme for video-and-language grounding tasks. To better encode the visual and linguistic modalities in the shared feature space, global level alignment loss focuses on training the network to capture context information, while the segment level alignment emphasizes fine-grained semantics. The application scenario of our multi-level alignment scheme is not restricted. It can be applied to video QA, video-text retrieval and video moment retrieval tasks, and was able to help the lightweight model achieve good performance on multiple datasets.

References

[1] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su, "Scene graph captioner: Image captioning based on structural visual representation," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 477–485, 2019.

[2] S. Lee, J.-W. Kim, Y. Oh, and J. H. Jeon, "Visual question answering over scene graph," in *2019 First International Conference on Graph Computing (GC)*. IEEE, 2019, pp. 45–50.

[3] W. Liang, F. Niu, A. Reganti, G. Thattai, and G. Tur, "Lrta: A transparent neural-symbolic reasoning framework with modular supervision for visual question answering," *arXiv preprint arXiv:2011.10731*, 2020.

[4] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[5] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8746–8755.

[6] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," *CVPR*, 2020.

[7] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9972–9981.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*, 2019.

[9] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[10] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.

[11] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+ language omni-representation pre-training," in *EMNLP*, 2020.

[12] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7331–7341.

[13] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[14] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[15] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SygXPaEYvH

[16] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7464–7473.

[17] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "Coot: Cooperative hierarchical transformer for video-text representation learning," *Advances in neural information processing systems*, vol. 33, pp. 22 605–22 618, 2020.

[18] X. Wang, L. Zhu, and Y. Yang, "T2vlad: global-local sequence alignment for text-video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5079–5088.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[22] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[23] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 336–11 344.

[24] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.

[25] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 162–11 173.

[26] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[27] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv–1807, 2018.

[28] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6210–6219.

[29] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[31] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[32] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[33] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[34] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[35] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[36] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo, "Tgif: A new dataset and benchmark on animated gif description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4641–4650.

[37] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4581–4591.

[38] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[39] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3537–3545.

[40] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2630–2640.

[41] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[42] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4631–4640.

[43] J. Lei, L. Yu, M. Bansal, and T. Berg, "Tvqa: Localized, compositional video question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1369–1379.

[44] Y. Jang, Y. Song, C. D. Kim, Y. Yu, Y. Kim, and G. Kim, "Video question answering with spatio-temporal reasoning," *International Journal of Computer Vision*, vol. 127, no. 10, pp. 1385–1412, 2019.

[45] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *AAAI*, 2019, pp. 9127–9134.

[46] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Agqa: A benchmark for compositional spatio-temporal reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 287–11 297.

[47] J. Liu, W. Chen, Y. Cheng, Z. Gan, L. Yu, Y. Yang, and J. Liu, "Violin: A large-scale dataset for video-and-language inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 900–10 910.

[48] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *European Conference on Computer Vision*. Springer, 2020, pp. 214–229.

[49] S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, and Z. Wang, "Hit: Hierarchical transformer with momentum contrast for video-text retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 915–11 925.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[51] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[52] L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang *et al.*, "Value: A multi-task benchmark for video-and-language understanding evaluation," in *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[54] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[56] P. H. Seo, A. Nagrani, and C. Schmid, "Look before you speak: Visually contextualized utterances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 877–16 887.

[57] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1686–1697.

[58] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[59] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 7–16.

[60] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," *arXiv preprint arXiv:1511.06361*, 2015.

[61] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[62] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1979–1988.