This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Adaptive Hierarchical Graph Reasoning with Semantic Coherence for Video-and-Language Inference

Juncheng Li[1] Siliang Tang[1]* Linchao Zhu[2] Haochen Shi[3] Xuanwen Huang[1]
Fei Wu[1] Yi Yang[1] Yueting Zhuang[1]
[1]Zhejiang University [2]ReLER, University of Technology Sydney [3]Université de Montréal
{junchengli, siliang, xwhuang, wufei, yangyics, yzhuang}@zju.edu.cn , linchao.zhu@uts.edu.au
haochen.shi@umontreal.ca

Abstract

Video-and-Language Inference is a recently proposed task for joint video-and-language understanding. This new task requires a model to draw inference on whether a natural language statement entails or contradicts a given video clip. In this paper, we study how to address three critical challenges for this task: judging the global correctness of the statement involved multiple semantic meanings, joint reasoning over video and subtitles, and modeling long-range relationships and complex social interactions. First, we propose an adaptive hierarchical graph network that achieves in-depth understanding of the video over complex interactions. Specifically, it performs joint reasoning over video and subtitles in three hierarchies, where the graph structure is adaptively adjusted according to the semantic structures of the statement. Secondly, we introduce semantic coherence learning to explicitly encourage the semantic coherence of the adaptive hierarchical graph network from three hierarchies. The semantic coherence learning can further improve the alignment between vision and linguistics, and the coherence across a sequence of video segments. Experimental results show that our method significantly outperforms the baseline by a large margin.

1. Introduction

Understanding video story involves analyzing and simulating human vision, language, thinking, and behavior, which is a significant challenge to current machine learning technology [21]. Recently, with the advances of large-scale video datasets [1, 5, 8, 24, 49], joint video-and-language understanding has received increased attention. Several video-and-language tasks have been proposed, such as video captioning [16, 48, 61, 14, 27, 13, 41, 50], text-to-video temporal grounding [15, 3, 6, 29, 36, 59, 63], and video question answering [28, 60, 47, 23, 25, 37]. In particular, Video-and-Language Inference (VLI) [33] is a recently proposed task to foster deeper investigations in video-and-language understanding. Given a video clip with aligned subtitles and a natural language statement based on the video content, a model needs to infer whether the statement entails or contradicts the given video clip. To support the study of this new task, a large-scale dataset, named VIOLIN (VIdeO-and-Language INference), is introduced.

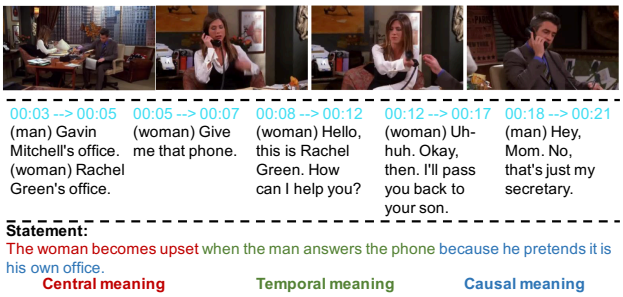

Figure 1: The first two rows show a video clip paired with its aligned subtitles. The third row shows a statement with multiple semantic meanings.

Compared with TVQA/video captioning where most QA pairs/captions focus on identifying explicit visual cues (*e.g.*, objects, actions, persons), VLI is more challenging and requires more sophisticated reasoning skills, such as interpreting human emotions and relations, understanding the events, and inferring causal relations of events throughout the video. First, a single statement may involve multiple semantic meanings, making it harder to judge the global correctness. As demonstrated in Figure 1, the statement consists of three semantic phrases. If the model recognizes the central meaning and the temporal meaning but ignores the

*Siliang Tang is the corresponding author.

causal meaning, it may make a wrong prediction. Secondly, VLI requires jointly reasoning over video and subtitles to achieve in-depth understanding of complex plots. To infer the causal meaning that the man pretends it is his own office, the model needs to jointly understand information from the video part and subtitle part. From the video part, the man and the woman are in the same office, and the man takes the phone from the woman. From the subtitle part, the man lies that the woman is his secretary. Only by combining the context from both the video and subtitle can the model further draw the inference. Thirdly, VLI requires reasoning for diverse interactions among characters and complex event dynamics over diverse scenarios. The VIOLIN dataset is collected from diverse sources to cover realistic visual scenes, including 5885 movie clips in addition to TV shows used in TVQA. The average clip length is 35.20s, while the length of most clips in TVQA is less than 15s.

In this paper, we propose a novel adaptive hierarchical graph reasoning with semantic coherence approach to overcome the aforementioned challenges. First, we introduce an adaptive graph construction mechanism to identify the multiple semantic meanings of the statement. This enables our approach to adaptively adjust the graph structure according to the semantic structures of the statement for the global correctness. Then, we present an adaptive hierarchical graph network (AHGN) to jointly reason over video and subtitles and model the complex social interactions. Specifically, we perform adaptive graph reasoning in three hierarchies: 1) segment-level reasoning, which achieves in-depth understanding of the video segments via utilizing the inherent alignment and complementary nature between visual frames and subtitles; 2) temporal-level reasoning, which models the long-range dependencies and diverse interactions between different segments to draw a global video understanding; 3) global-level reasoning, which judges the global correctness of the statement by incorporating the inferences from different reasoning steps.

Furthermore, the semantic coherence throughout AHGN is crucial to achieving global understanding of the video. Therefore, we introduce a novel semantic coherence learning (SCL) method to encourage the cross-modal semantic coherence at the segment level and the cross-level semantic coherence between temporal level and global level. Specifically, the semantic coherence learning contains two regularization terms: an optimal transport distance term that measures the cross-modality alignment between the visual nodes and subtitle nodes, and a mutual information term that evaluates the semantic coherence between the temporal nodes and global nodes.

The experiments show that our approach significantly outperforms the baselines by a large margin, and further ablation study demonstrates the effectiveness of each component. In summary, our contributions are mainly three folds:

- We propose a novel adaptive hierarchical graph network (AHGN) that performs joint reasoning over video and subtitles in three hierarchies, where the graph reasoning structure is adaptively adjusted according to the semantic structures of the statement.
- Our semantic coherence learning (SCL) method improves the alignment between video and subtitles, and the coherence across a sequence of video segments.
- Extensive experiments show that our method significantly outperforms the baseline by a large margin.

2. Related Work

Visual Entailment Given a natural image premise and a natural language hypothesis, the goal of visual entailment (VE) [51] is to predict whether the image semantically entails the text. To realize this task, the SNLI-VE dataset is built based on the Stanford Natural Language Inference corpus and Flickr30k dataset [54]. Also, Suhr *et al.* [46] propose a similar task to determine whether a natural language caption is true about a photograph. In contrast to the visual entailment, which is limited to a static image, video-and-language inference involves complex temporal dynamics and requires the model to understand the relationship between different visual scenes to draw the inference. In this paper, we propose an approach to model the complex inter- and intra- modality interactions and further infer in-depth rationale in three hierarchies.

Video-and-Language Research Recent years have witnessed the flourishing development in vision-and-language research [4, 15, 56, 31, 30, 17]. Several large-scale video datasets [1, 5, 8, 24, 49] and video-and-language tasks have been proposed, such as video captioning [16, 44, 57, 58, 48, 52, 27, 13, 41, 50], text-based video moment retrieval [15, 3, 6, 29, 36], and video question answering [28, 62, 47, 43, 25, 37]. Video caption is the task of generating text descriptions from video input, text-based video moment retrieval requires localizing video segments from natural language queries, and video question answering is aimed 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, which hardly incorporate story-level understanding. In contrast, video-and-language inference [33] requires not only explicit visual cues but also more sophisticated reasoning skills, such as inferring reasons and interpreting human emotions. These abilities can be used to detect anomalous intent from surveillance and discriminatory or antisocial contents from online videos, which are usually expressed implicitly. Similar to [28] for TVQA, the baseline model for VLI [33] utilizes multi-stream neural network [33, 28] with bidirectional attention [28, 42, 55] to interact the statement

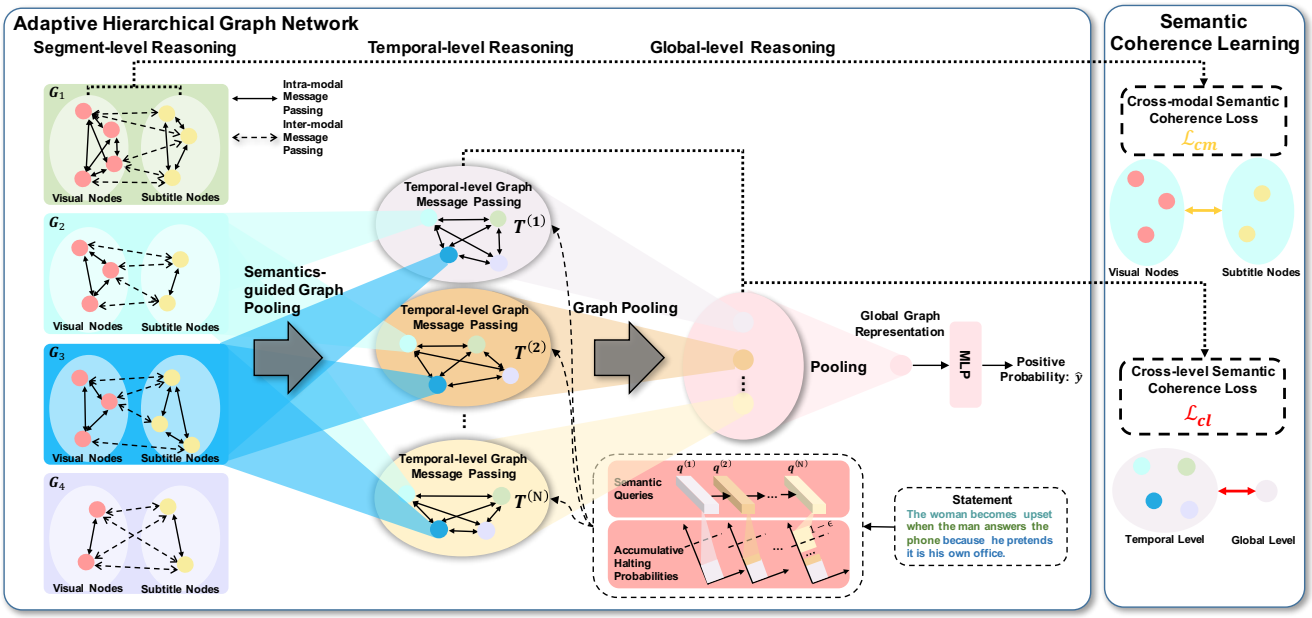

Figure 2: Overview of the proposed framework. Each sub-graph G_i is constructed with subtitle word nodes $\{s_i\}$ and time-aligned visual frame nodes $\{v_i\}$. Our AHGN performs joint reasoning over video and subtitles in three hierarchies. $T^{(i)}$ represents a temporal-level sub-graph, which is adaptively constructed according to the semantic structures of the statement. The semantic coherence learning explicitly promotes the cross-modal and cross-level semantic coherence of AHGN.

with subtitles and visual frames separately, and then fuse the independent attention representations to draw the final inference. However, in our view, there are two main limitations of this type of approaches: 1) it fails to leverage the temporal alignment and complementary nature between visual frames and subtitles, which is a crucial step to achieve in-depth understanding of the video; and 2) it equally uses each word in the statement to attend every visual frame and subtitle word, and then do single-step classification, without considering explicit semantic structures. Our work instead models the semantic alignment between visual frames and subtitles, and the graph structure is adaptively adjusted according to the semantics of the statement.

3. Method

As described in Figure 2, the adaptive hierarchical graph reasoning framework mainly consists of two components: 1) adaptive hierarchical graph network (Sec 3.1), and 2) semantic coherence learning (Sec 3.2). Given a video clip, paired with its aligned subtitles, and a natural language statement, the adaptive hierarchical graph network (AHGN) performs reasoning in three hierarchies, of which the graph structure is adaptively adjusted according to the semantics of the statement. Finally, the prediction layer performs classification using the global graph representation. The semantic coherence learning (SCL) is further introduced to explicitly promote the cross-modal and cross-level semantic coherence of the adaptive hierarchical graph network.

3.1. Adaptive Hierarchical Graph Network

3.1.1 Graph Construction

For the given video clip, we extract the visual features using ResNet101 [20] trained on ImageNet [11] and apply a single-layer MLP to obtain the visual node representations $V = \{v_i\}_{i=1}^{l_v}$. For its associated subtitles, we tokenize them into a word sequence and employ a pre-trained BERT [12] encoder, followed by a single-layer MLP, to obtain the subtitle word node representations $S = \{s_i\}_{i=1}^{l_s}$. Each node v_i corresponds to a visual frame, and each node s_i corresponds to a subtitle word. We then align each subtitle sentence $S_i = \{s_i\}_{i=sst_i}^{sst_i+L}$ with a sequence of visual frames $V_i = \{v_i\}_{i=vst_i}^{vst_i+K}$ whose timestamps overlap with the subtitle timestamp (sst_i is the start index of subtitle nodes in S_i, L is the number of subtitle nodes in S_i, similar denotation for V_i). The frame-subtitle pair $< S_i, V_i >$ corresponds to a semantic segment of the video and makes up a cross-modal sub-graph G_i, where we perform segment-level reasoning. Thus, we split the original video-level graph into a sequence of sub-graphs (*i.e.* $G = \{G_i\}_{i=1}^{M}$).

3.1.2 Segment-level Reasoning

To utilize the inherent alignment and complementary nature between visual frames and subtitles, for each segment-level sub-graph, our adaptive hierarchical graph network first models the inter- and intra- modality interactions, using the gated inter-modal message passing (GER) and the gated intra-modal message passing (GRA). Then, we summarize

the semantic information of each sub-graph into a temporal-level node representation by semantics-guided graph pooling under the guidance of the semantic query.

Gated Inter-modal Message Passing Here we take the message passing from subtitle nodes to visual nodes as an illustration. Given segment-level sub-graph $G_i =< S_i, V_i >$, we first learn the cross-modal adjacency correlation matrix $A \in \mathbb{R}^{K \times L}$ by calculating the similarity of each pair of visual node v_i and subtitle node s_j as:

$$A = V_i^T \cdot S_i, \quad a_{ij} = v_i^T \cdot s_j \tag{1}$$

Then we compute the message m_i^v that denotes the message from subtitle nodes to visual node v_i: $m_i^v = \sum_j^L a_{ij} \cdot s_j$. Next, we get the visual guidance g_v and subtitle guidance g_s by performing average pooling on the visual nodes and subtitle nodes respectively. The context gate c_i^v is further determined as:

$$c_i^v = \sigma(W_1[g_v, v_i, g_s] + b_1) \tag{2}$$

where $W_1 \in \mathbb{R}^{d \times 3d}$, $b_1 \in \mathbb{R}^{d \times 1}$, and $\sigma(\cdot)$ denotes the sigmoid function. The context gate c_i^v controls the flow of linguistic information from subtitles to vision:

$$\tilde{v_i} = (1 - c_i^v) \odot v_i + c_i^v \odot m_i^v \tag{3}$$

where $\odot$ denotes the Hadamard product. As a consequence, $\tilde{v_i} \in \mathbb{R}^{d \times 1}$ represents the linguistic-refined visual node, and we can obtain the visual-refined subtitle node $\tilde{s_i} \in \mathbb{R}^{d \times 1}$ in a similar manner but reversed order, and $\tilde{G_i} =< \tilde{S_i}, \tilde{V_i} >$ represents the sub-graph after GER.

Gated Intra-modal Message Passing For $\tilde{V_i}$ and $\tilde{S_i}$, we further refine them with intra-modal local context information using GRA. GRA is similar to the GER but models the intra-modal relations.

The GRA first computes the intra-modal adjacency correlation matrixes $A^v \in \mathbb{R}^{K \times K}$ and $A^s \in \mathbb{R}^{L \times L}$ by calculating node-wise similarity, and then get the aggregated messages n_i^v and n_i^s according to the weight matrixes. Next, the GRA computes the context gates for visual/subtitle nodes based on the visual/subtitle guidance and corresponding visual/subtitle node representations respectively. Finally, GRA updates the visual nodes and subtitle nodes respectively, controlled by the context gates. As a consequence, $\hat{V_i}$ and $\hat{S_i}$ represent refined nodes of G_i after inter- and intra-modality reasoning.

Semantics-guided Graph Pooling After obtaining segment-level refined node representations $\hat{V_i} \in \mathbb{R}^{d \times K}$ and $\hat{S_i} \in \mathbb{R}^{d \times L}$ using inter- and intra- modality local context, we further aggregate $G_i =< \hat{V_i}, \hat{S_i} >$ to a temporal-level node representation by semantics-guided graph pooling. We first extract a semantic query $q^{(n)} \in \mathbb{R}^{d \times 1}$ from the statement using attentive aggregation. Then, we employ the semantic query to attend each visual node and subtitle node independently, and then use the attention weights to aggregate $\hat{V_i}$ and $\hat{S_i}$:

$$C_i^{(n)} = softmax(\hat{V_i}^T(W_2 q^{(n)})), \quad v_i^{(n)} = \hat{V_i} C_i^{(n)} \tag{4}$$

And we can obtain $s_i^{(n)} \in \mathbb{R}^{d \times 1}$ in a similar manner. Then, we compute the context gate $\gamma_i^{(n)}$ based on the visual guidance, subtitle guidance, and the semantic query, which controls the fusion of visual and linguistic information:

$$\gamma_i^{(n)} = \sigma(W_3[g_v, q^{(n)}, g_s] + b_3) \tag{5}$$

$$t_i^{(n)} = (1 - \gamma_i^{(n)}) \odot v_i^{(n)} + \gamma_i^{(n)} \odot s_i^{(n)} \tag{6}$$

where $W_3 \in \mathbb{R}^{d \times 3d}$ and $b_3 \in \mathbb{R}^{d \times 1}$. As a consequence, we pool each segment-level sub-graph G_i to a temporal-level node $t_i^{(n)} \in \mathbb{R}^{d \times 1}$ and obtain temporal-level sub-graph $T^{(n)} \in \mathbb{R}^{d \times M}$, guided by the semantic query $q^{(n)}$.

3.1.3 Temporal-level Reasoning

The video clips are collected from TV shows and movies, which contain complex event dynamics and diverse character interactions across multiple segments. Therefore, segment-level reasoning is not sufficient, and we present temporal-level reasoning to model long-range relationships among multiple video segments to draw a global understanding. In this section, we first introduce how to construct multiple temporal-level sub-graphs adaptively, and then present how to perform reasoning on them.

Adaptive Temporal-level Sub-graph Construction Different statements may have semantic structures of varying complexity. For more complex statements, it is reasonable to construct more temporal-level sub-graphs, which focus on different semantic parts of the statement. Thus, we introduce the adaptive temporal-level sub-graph construction to adaptively adjust the number of temporal-level sub-graphs. Concretely, we extract a variable number of semantic queries from the statement, and use them to perform semantics-guided graph pooling respectively to construct multiple temporal-level sub-graphs. The semantic queries provide guidance to their corresponding temporal-level sub-graphs about which semantic parts they should focus on.

Given l_h-word statement $H = \{h_i\}_{i=1}^{l_h}$, we extracts N semantic queries $\{q^{(n)}\}_{n=1}^{N}$. At each step n, we first compute the attention weights $R^{(n)} \in \mathbb{R}^{l_h \times 1}$ based on the previous semantic query $q^{(n-1)}$ and the sentence-level embedding of the statement $g_h \in \mathbb{R}^{d \times 1}$ obtained by performing average pooling on the statement words H, given by:

$$R^{(n)} = softmax(H^T(W_r[g_h, q^{(n-1)}])) \tag{7}$$

Then, we obtain the semantic query $q^{(n)}$ using the attention weights to summarize the statement words: $q^{(n)} = HR^{(n)}$.

To adaptively determine how many temporal-level subgraphs are supposed to be constructed, we introduce a self-halting mechanism that outputs the probability of stopping generating more queries: $h^{(n)} = \sigma(W_h q^{(n)} + b_h)$ where $W_h \in \mathbb{R}^{d\times d}$, $b_h \in \mathbb{R}^{d\times 1}$, and the accumulative halting probability is further determined as: $P^{(n)} = \sum_{i=1}^{n} h^{(i)}$. When the accumulative halting probability exceeds a threshold $1-\epsilon$ or n reaches the pre-defined maximum value N_{max}, the process will stop. To promote the generation efficiency, we define a loss term based on the final number of queries:

$$\mathcal{L}_{qe} = \tau N \tag{8}$$

where τ is the query efficiency hyper-parameter.

The message passing at temporal-level graph is similar to the GRA. Given $T^{(n)}$, we first compute the adjacency correlation matrix $E^{(n)} \in \mathbb{R}^{M\times M}$, then sum up the incoming messages, and finally update the node representations using gate mechanism. After obtaining refined $\tilde{T}^{(n)}$, we further use semantic query $q^{(n)}$ to pool $\tilde{T}^{(n)}$ to a global semantic representation o_n:

$$U^{(n)} = softmax((\tilde{T}^{(n)})^T(W_4 q^{(n)})), \quad o_n = \tilde{T}^{(n)} U^{(n)} \tag{9}$$

where $W_4 \in \mathbb{R}^{d\times d}$, $U^{(n)} \in \mathbb{R}^{K\times 1}$ is the attention weight vector, and $o_n \in \mathbb{R}^{d\times 1}$ is the global node representation based on semantic query $q^{(n)}$.

3.1.4 Global-level Reasoning

After temporal-level reasoning, we obtain a set of N global semantic representations $\{o_i\}_{i=1}^{N}$. We perform average pooling on the set of global semantic representations to generate a d-dimensional global graph representation capturing whole semantics of video and statement. Finally, the global graph representation is passed through the MLP with a sigmoid activation to predict the probability of the input statement being positive.

3.2. Semantic Coherence Learning

We specify the semantic coherence learning of AHGN as the cross-modal semantic coherence among segment-level nodes and the cross-level semantic coherence across temporal-level nodes and global-level nodes.

3.2.1 Cross-modal Semantic Coherence

To achieve in-depth understanding of the semantics in video and subtitles, a model needs to align the inter-modality semantics between visual nodes and subtitle nodes. However, most of the previous methods seek advanced attention mechanisms to simulate soft alignment, with no training signals to explicitly encourage alignment. Differently, we leverage recent advances in Optimal Transport (OT) [7] to encourage the cross-modal semantic coherence of each sub-graph, which can further refine the segment-level reasoning on every sub-graph G_i.

Optimal transport evaluates the correspondence between two distributions. OT-based learning aims to optimize distribution matching via minimizing the cost of transporting one distribution to another, which provides explicit signals to minimizing the embedding distance between the modalities. Recently, it has been explored in some fields. Liu *et al.* [34] model semantic correspondence as an optimal transport problem. Su *et al.* [45] apply optimal transport to 3D shape matching and comparison. Chen *et al.* [7] solve cross-domain alignment by minimizing the optimal transport plan between domains. Here, we adopt OT to refine the segment-level reasoning on sub-graph G_i. By optimizing the node distance and edge distance between visual nodes and subtitle nodes, we further foster the semantic coherence learning of gated inter- and intra- modal message passing.

Specifically, we adapt Wasserstein distance (WD) [35] for node matching and Gromov-Wasserstein distance (GWD) [40, 9] for edge matching. We define two distributions $\mu_s \in P(S), \mu_v \in P(V)$ as: $\mu_s = \sum_{i=1}^{n} p_i^s \delta_{s_i}$ and $\mu_v = \sum_{j=1}^{m} p_j^v \delta_{v_i}$ where δ_{s_i} denotes the Dirac function centered on s_i. Without ambiguity, we reuse m and n to represent the number of visual nodes and subtitle nodes for simplicity. $\Pi(\mu_s, \mu_v)$ denotes all the joint distributions, with marginals $\mu_s(s)$ and $\mu_v(v)$. Let $p^s = \{p_i^s\}_{i=1}^{n} \in \Delta_n$ and $p^v = \{p_j^v\}_{j=1}^{m} \in \Delta_m$ denote the $n-$ and $m-$ dimensional weight vectors respectively, where $\sum_{i=1}^{n} p_i^s = \sum_{j=1}^{m} p_j^v = 1$, and both p^s and p^v are probability distributions. $\Pi(p^s, p^v) = \{T \in \mathbb{R}^{n\times m} | T1_m = p^s, T^T 1_n = p^v\}$, where T denotes the transport plan and T_{ij} represents the amount of mass shifted from p_i^s to p_j^v. Formally, the optimal transport distance is defined as:

$$\mathcal{D}(\mu_s, \mu_v) = \inf_{\gamma\in\Pi(\mu_s,\mu_v)} \mathbb{E}_{(s,v)\sim\gamma,(s',v')\sim\gamma}[c(s,v) + \mathcal{L}(s,v,s',v')] \tag{10}$$

$$= \min_{T\in\Pi(p^s,p^v)} \sum_{i,i',j,j'} T_{ij}[\lambda c(s_i, v_i) + T_{i',j'} \mathcal{L}(s_i, v_j, s_{i'}, v_{j'})] \tag{11}$$

where λ is the weight hyper-parameter, $c(s_i, v_j)$ is the cost function that evaluates the node similarity between s_i and v_j using cosine distance, and $\mathcal{L}(s_i, v_j, s_{i'}, v_{j'}) = ||c_1(s_i, s_{i'}) - c_2(v_j, v_{j'})||$ is the cost function evaluating the similarity between two pairs of nodes $(s_i, s_{i'})$ and $(v_j, v_{j'})$.

We apply the Sinkhorn algorithm [10, 39] to obtain the optimal transport distance $\mathcal{D}(\mu, \nu)$, following [7, 2]. Then, the calculated optimal transport distance is used as the cross-modal semantic coherence loss:

$$\mathcal{L}_{cm} = \alpha \mathcal{D}(\mu, \nu) \tag{12}$$

where α is the weight hyper-parameter. $\mathcal{L}_{cm}$ provides an explicit training objective to encourage semantic alignment of each sub-graph.

3.2.2 Cross-level Semantic Coherence

Every video clip is composed of a sequence of segments, which follow a consistent theme intrinsically and narrate the event coherently. Motivated by this fact, we present the cross-level semantic coherence to promote the semantic coherence across temporal-level nodes and global-level nodes. By keeping the cross-level semantic coherence, we can greatly improve the graph representation's quality for story-level understanding.

Mutual information have been widely utilized for representation learning (*e.g.* variational autoencoder [26], beta-VAE [22]), while we are the first to exploit mutual information to learn semantic coherence of the video. For temporal-level nodes $\{t_i^{(n)}\}_{i=1}^{M}$ and its corresponding global node o_n, we maximize the average mutual information between them as: $\frac{1}{M\times N}\sum_{n=1}^{N}\sum_{i=1}^{M}\mathcal{I}(t_i^{(n)};o_n)$. The mutual information maximization procedure can encourage segment-level reasoning and temporal-level reasoning to encode more underlying semantic information that is coherent in the video. To compute the mutual information, we use Noise-Contrastive Estimation (NCE) [19, 18, 38] to estimate it as follows:

$$\hat{\mathcal{I}}(t_i^{(n)};o_n) := \mathbb{E}_{\mathbb{P}}[T_\varphi(t_i^{(n)},o_n) - \mathbb{E}_{\tilde{\mathbb{P}}}[log\sum_{t_i^{(n)\prime}} e^{T_\varphi(t_i^{(n)\prime},o_n)}]] \tag{13}$$

where T_φ is a discriminator modeled by a neural network with parameters φ, $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are the distribution of $t_i^{(n)}$. Then the cross-level semantic coherence can be formulated as:

$$\mathcal{L}_{cl} = -\beta\frac{1}{M\times N}\sum_{n=1}^{N}\sum_{i=1}^{M}\hat{\mathcal{I}}(t_i^{(n)};o_n) \tag{14}$$

where β is weight hyper-parameter. We add this term to the loss function and optimize θ and φ simultaneously. Thus, the total loss is given by:

$$\mathcal{L} = \mathcal{L}_{ent} + \mathcal{L}_{qe} + \mathcal{L}_{cm} + \mathcal{L}_{cl} \tag{15}$$

where $\mathcal{L}_{ent}$ is the original cross-entropy loss.

4. Experiments

4.1. Experimental Setup

Dataset The VIOLIN dataset contains 15887 video clips collected from 4 popular TV shows and movie clips from YouTube channels covering thousands of movies. The average length of each video clip is 35.20s with 3 frames per second and each statement has 18 words on average. Each video clip is annotated with 3 pairs of positive/negative statements, resulting in 95322 (V, S, H) triplets in total. It is divided into 76122, 9600, and 9600 triplets for training, validation, and testing, respectively. Model performance is evaluated via binary classification accuracy.

	Method	Vision	Text	Accuracy
1	MTS	Img	GloVe	60.33
2	MTS	Img	BERT	67.60
3	MTS	C3D	BERT	67.23
4	MTS	Det	BERT	67.84
5	DIFFPOOL-Split	Img	BERT	59.46
6	DIFFPOOL-Whole	Img	BERT	56.43
7	XML	Img	BERT	66.32
8	HERO (pre-trained)	Img	BERT	68.59
9	**Ours- AHGN + SCL**	Img	BERT	**71.38**

Table 1: Quantitative results on the VIOLIN dataset.

Implementation Details The dimension of the node embedding is set to 512. For the visual frames that are not paired with any subtitles, we assign it to the neighboring frame-subtitle pair. We set the halting threshold ϵ to 0.1, the maximum query number N_{max} to 5, and the query efficiency τ to 0.05. In the training stage, the learning rate is $1e^{-4}$ and the batch size is 128.

Baselines 1-4) MTS: these baselines are based on the multi-stream architecture (please refer to Sec 2 for details). We also compare our model with the state-of-the-art hierarchical graph representation learning method and hierarchical transformer-based models for video modeling: 5,6) DIFFPOOL: a differentiable graph pooling module [53] that learns soft cluster assignment matrix for nodes in each layer. We try two versions: DIFFPOOL-whole that uniformly regards each video frame and subtitle word as the node, and DIFFPOOL-split that constructs a video frame graph and a subtitle word graph respectively. 7) XML: Cross-modal Moment Localization (XML) modular network [29] is a recently proposed transformer-based method for TV show retrieval. 8) HERO: a transformer-based framework [32] for video-and-language pre-training. It has two standard hierarchies with fixed structures for local and global context computation.

4.2. Results

We summarize the results in Table 1, where our method significantly outperforms all baselines. Compared with the original baseline presented in [33], our method surpasses it by 6.69% relatively on accuracy. The comparison with DIFFPOOL indicates the effectiveness of our semantics-guided graph pooling scheme. The semantics-guided graph pooling can better control the complementary information fusion of subtitles and visual frames. Furthermore, XML models the relationships directly on the whole sequences of

	Method	Vision	Text	Accuracy
1	MTS	Img	BERT	67.60
2	AHGN	Img	BERT	69.76
3	AHGN + $\mathcal{L}_{cm}$	Img	BERT	70.19
4	AHGN + $\mathcal{L}_{cl}$	Img	BERT	70.47
5	AHGN + $\mathcal{L}_{cm}$ + $\mathcal{L}_{cl}$	Img	BERT	**71.38**

Table 2: Main ablation results.

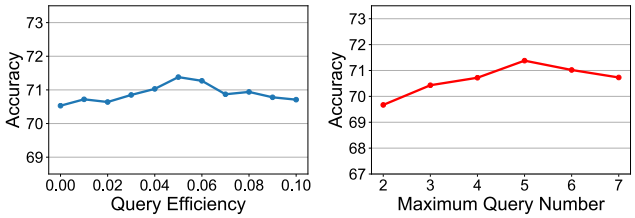

Figure 3: Ablation across with respect to τ and N_{max}.

frames and subtitles, while we model the interaction in three hierarchies, leveraging the temporal alignment and complementary nature. The results show the efficiency of our hierarchical strategy. HERO first fuses every frame-subtitle pairs using a cross-modal transformer and then applies temporal transformer on the merged sequence of them. Our method differs as it represents the context at three different levels of granularity with adapative graph structure, which outperforms the pre-trained HERO by 2.79%.

4.3. In-Depth Analysis

Effectiveness of Individual Components We conduct an ablation study to illustrate the effectiveness of each component in Table 2. Comparing MTS and AHGN (Row 1 vs Row 2), AHGN significantly contributes 2.16% to the improvement on accuracy. The results of Row 3 and Row 4 validate the superiority of the cross-modal and cross-level semantic coherence, respectively. Meanwhile, the results indicate that the introduced two losses can promote the cross-modal and cross-level semantic coherence of the AHGN in a mutually rewarding way. Finally, the semantic coherence learning (Row 5) takes up 2.32% of the relative gain on accuracy.

Analysis on AHGN We further perform in-depth analysis for the adaptive hierarchical graph network. We investigate the contribution of the proposed three graph operations and the adaptive graph structure. We start with the backbone model, which keeps the adaptive hierarchical structure but removes the proposed graph operations. For the ablation model with fixed graph structure, we set the number of temporal-level sub-graphs as a hyper-parameter. Specifically, we extract fixed number of semantic queries from the statement and use them to construct temporal-level sub-graphs respectively. We test different number of temporal-level sub-graphs and report the best performance when the number of temporal-level sub-graph is 3. Note that the detailed results are provided in the supplementary material. For the ablation model without temporal-level reasoning, we directly perform attentive pooling using semantic queries on the nodes after segment-level reasoning.

	Method	Graph Operation			Graph Structure		Accuracy
		GER	GRA	Temp	fixed	adaptive	
1	Backbone (fixed)				✓		64.65
2	Backbone					✓	65.32
3	+ GER	✓				✓	67.07
4	+ GRA		✓			✓	66.74
5	+ Temporal			✓		✓	66.93
6	+ GER + GRA	✓	✓			✓	68.03
7	+ GER + Temp	✓		✓		✓	68.64
8	+ GRA + Temp		✓	✓		✓	68.15
9	AHGN	✓	✓	✓		✓	69.76
10	AHGN (fixed)	✓	✓	✓	✓		68.91

Table 3: We conduct comparison by varying the individual components of the AHGN.

Table 3 summarizes the results, which indicate the following. First, the adaptive graph structure is more effective than the fixed graph structure, which enables our AHGN to dynamically adjust the graph structure according to the statement. Second, the inter- and intra- modality reasoning at segment-level significantly improve the performance by modeling the inherent alignment and complementary nature between visual frames and subtitles. Third, the temporal-level reasoning is a crucial step to achieve in-depth understanding of the video.

Ablation of the Adaptive Graph Construction We explore the impact of the query efficiency hyper-parameter τ and the maximum query number N_{max} for the adaptive graph construction (AGC). The higher τ means that the tolerance for more queries is lower. As illustrated in Figure 3, the performance keeps increasing when the τ is increased from 0.02 to 0.05. When we continue to increase the τ, the performance decreases because too large τ limits the number of semantic queries to one or two. Further, the maximum query number N_{max} of 5 and 6 are enough and can generally provide good performance while higher N_{max} does harm to the performance and reduces the efficiency.

4.4. Qualitative Analysis

For a more intuitive view of how our model works for the VLI task, we visualize two qualitative examples in Figure 4. The attention weights of semantic queries reflect the semantic parts which their corresponding temporal-level sub-graphs focused on. As shown in Figure 4, different temporal-level sub-graphs focus on different semantic phrases of the statement and pay attention to the video segments that are most related to their semantics.

Visualization Fig 5 visualizes the inter- and intra- modality gate values (top row). Some key frames are in the bottom row. The $t = 1 \sim 3$ frames show that a man talks to someone angrily, and the inter-modality gates are acti-

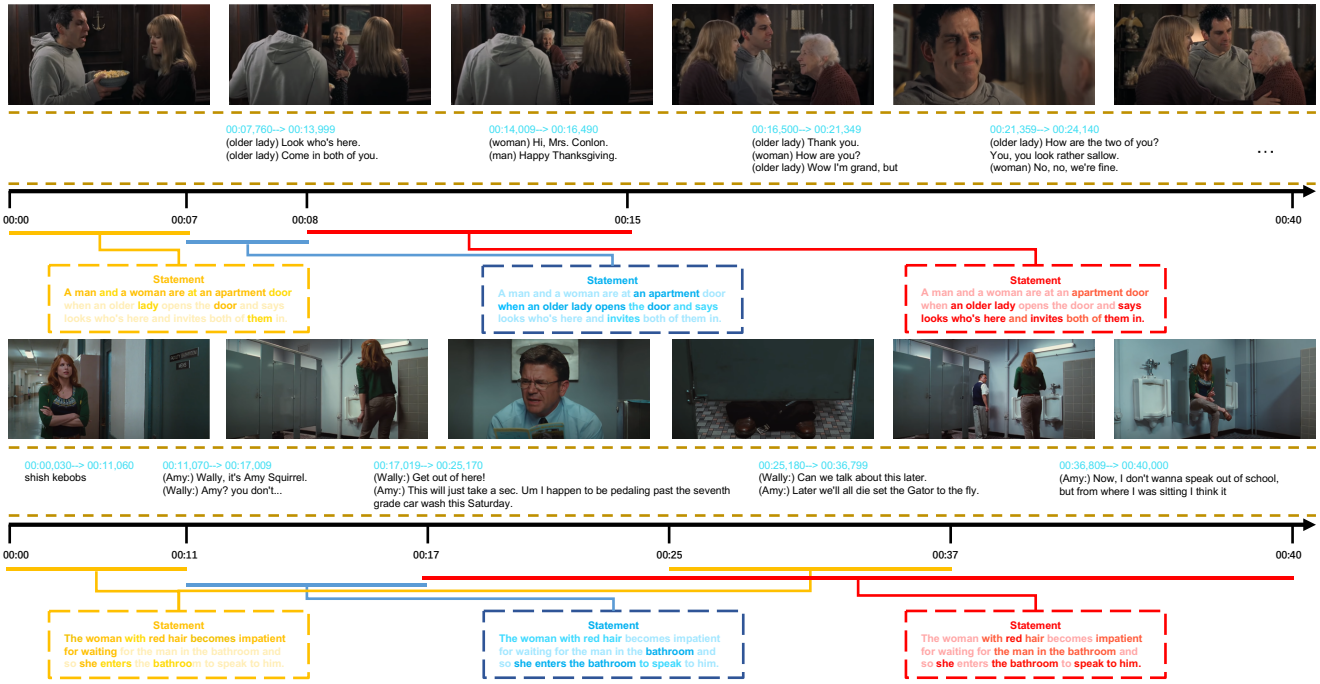

Figure 4: Qualitative Examples. The attention weights of different semantic queries are illustrated by the depth of color, and their corresponding video moments with high attention weights during semantics-guided graph pooling are also shown.

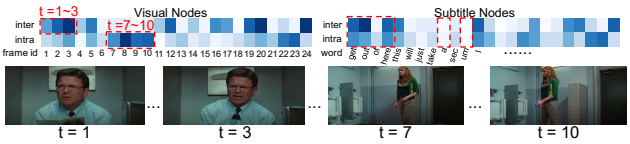

Figure 5: Visualization of the GER and GRA.

vated to combine linguistic context from subtitles. This helps the model to understand what he says and why he is angry. For the $t = 7 \sim 10$ frames, the intra-modality gates are well activated to combine temporal information from local visual context. Thereby, the model can infer that the woman is coming in instead of standing still. Also, gate values of words "get" and "here" are both high. It indicates that GER and GRA cooperate with each other to infer the complete semantics and visual context. Some words like "a" and "um" receive low gate values. Fig 6 provides visualization on temporal-level node feature space and learned cross-modality alignment matrices between visual and subtitle nodes. We observe that **with** $\mathcal{L}_{cl}$ (Fig 6.b), the nodes from the same temporal-level sub-graph (nodes of the same color) tend to be more tightly related, compared with the features trained **without** $\mathcal{L}_{cl}$ (Fig 6.a). **Without** $\mathcal{L}_{cm}$ (Fig 6.c), the learned alignment matrix is much denser and noisier than the alignment matrix learned **with** $\mathcal{L}_{cm}$ (Fig 6.d).

5. Conclusions

In this paper, we introduce an adaptive hierarchical graph reasoning with semantic coherence approach for Video-and-Language Inference. Our adaptive hierarchical graph network performs in-depth reasoning over video frames and subtitles in three hierarchies, where the graph reasoning structure is adaptively determined by the semantic structures of the statement. Further, we present a semantic coherence learning algorithm to encourage the cross-modal and cross-level semantic coherence of the adaptive hierarchical graph network. The experimental results show that our method outperforms the baselines by a large margin.

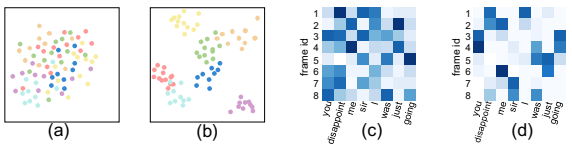

Figure 6: (a)(b): t-SNE visualization of temporal-level nodes without/with $\mathcal{L}_{cl}$. (c)(d): alignment matrixes without/with $\mathcal{L}_{cm}$.

Acknowledgment

This work has been supported in part by NSFC (No. 61625107), Zhejiang NSF (LR21F020004), National Key Research and Development Program of China (2018AAA0101900), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Zhejiang University iFLYTEK Joint Research Center, Chinese Knowledge Center of Engineering Science and Technology (CKCEST).

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

[2] David Alvarez-Melis and Tommi S Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 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, 2015.

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

[7] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Liu Jingjing. Graph optimal transport for cross-domain alignment. *arXiv preprint arXiv:2006.14744*, 2020.

[8] Yu Cheng, Quanfu Fan, Sharath Pankanti, and Alok Choudhary. Temporal sequence modeling for video event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2227–2234, 2014.

[9] Samir Chowdhury and Facundo Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.

[14] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017.

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

[16] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2712–2719, 2013.

[17] Jiannan Guo, Haochen Shi, Yangyang Kang, Kun Kuang, Siliang Tang, Zhuoren Jiang, Changlong Sun, Fei Wu, and Yueting Zhuang. Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[18] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[19] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 13(1):307–361, 2012.

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

[21] Yu-Jung Heo, Kyoung-Woon On, Seongho Choi, Jaeseo Lim, Jinah Kim, Jeh-Kwang Ryu, Byung-Chull Bae, and Byoung-Tak Zhang. Constructing hierarchical q&a datasets for video story understanding. *arXiv preprint arXiv:1904.00623*, 2019.

[22] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[23] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[25] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*, 2017.

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[28] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[29] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. *arXiv preprint arXiv:2001.09099*, 2020.

[30] Juncheng Li, Siliang Tang, Fei Wu, and Yueting Zhuang. Walking with mind: Mental imagery enhanced embodied qa. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1211–1219, 2019.

[31] Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, and William Yang Wang. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12123–12132, 2020.

[32] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[33] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020.

[34] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020.

[35] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 5859–5870, 2018.

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

[37] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017.

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[39] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[40] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

[41] Yunchen Pu, Martin Renqiang Min, Zhe Gan, and Lawrence Carin. Adaptive feature abstraction for translating video to text. *arXiv preprint arXiv:1611.07837*, 2016.

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

[43] Kai Shen, Lingfei Wu, Siliang Tang, Fangli Xu, Zhu Zhang, Yu Qiang, and Yueting Zhuang. Ask question with double hints: Visual question generation with answer-awareness and region-reference, 2021.

[44] Kai Shen, Lingfei Wu, Fangli Xu, Siliang Tang, Jun Xiao, and Yueting Zhuang. Hierarchical attention based spatial-temporal graph-to-sequence learning for grounded video description. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 941–947. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[45] Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259, 2015.

[46] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[47] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[48] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[49] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2295–2304, 2016.

[50] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4581–4591, 2019.

[51] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[52] Jun Xu, Tao Mei, Ting Yao, and Yong 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*, pages 5288–5296, 2016.

[53] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances*

in neural information processing systems, pages 4800–4810, 2018.

[54] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[55] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

[56] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020.

[57] Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. Poet: Product-oriented video captioner for e-commerce. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1292–1301, 2020.

[58] Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Tan Jiang, Jingren Zhou, Hongxia Yang, and Fei Wu. Comprehensive information integration modeling framework for video titling. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2744–2754, 2020.

[59] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *Proc 35 AAAI Conf on Artificial Intelligence*, 2021.

[60] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4):1032–1041, 2019.

[61] Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. Relational graph learning for grounded video description generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3807–3828, 2020.

[62] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017.

[63] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.