GHAN: Graph-Based Hierarchical Aggregation Network for Text-Video Retrieval

Yahan Yu[1*], **Bojie Hu**[1] **and Yu Li**[1,2]
[1] Tencent Minority-Mandarin Translation, Beijing, China
[2] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
jasmineyuyh@gmail.com, bojiehu@tencent.com, 20112043@bjtu.com

Abstract

Text-video retrieval focuses on two aspects: cross-modality interaction and video-language encoding. Currently, the mainstream approach is to train a joint embedding space for multi-modal interactions. However, there are structural and semantic differences between text and video, making this approach challenging for fine-grained understanding. In order to solve this, we propose an end-to-end graph-based hierarchical aggregation network for text-video retrieval according to the hierarchy possessed by text and video. We design a token-level weighted network to refine intra-modality representations and construct a graph-based message passing attention network for global-local alignment across modality. We conduct experiments on the public datasets MSR-VTT-9K, MSR-VTT-7K and MSVD, and achieve Recall@1 of 73.0%, 65.6%, and 64.0% , which is 25.7%, 16.5%, and 14.2% better than the current state-of-the-art model.

1 Introduction

Text-Video Retrieval (TVR) is a fundamental research task in multimodal video and language understanding, which aims to retrieve the most relevant video for a given text query, or to retrieve relevant text for a given video query. The mainstream modeling approaches for TVR are to jointly learn cross-modal interactive information between video and text in the same representation space (Lei et al., 2021; Dzabraev et al., 2021; Luo et al., 2020; Liu et al., 2021; Cheng et al., 2021).

However, the mismatched problem of information capacity and information density between video and text is not fully studied in these approaches. As shown in Figure 1, there are structural and semantic differences between the modalities, making this approach challenging for TVR. The video itself expresses a much wider range of global representations than the text, so the textual semantic usually cannot be fully mapped to every detailed information of the video. If the content mentioned by the text appears in a small number of frames in video, the text may be misled by unrelated semantics when interacting across modalities. In addition, the local representations in text are usually more specific than that in video. Due to different annotators, text may have different description habits and contain different semantic ranges, such as using *person/man/actor* to describe a character appearing in the video, which makes it difficult with semantic understanding and alignment across modalities.

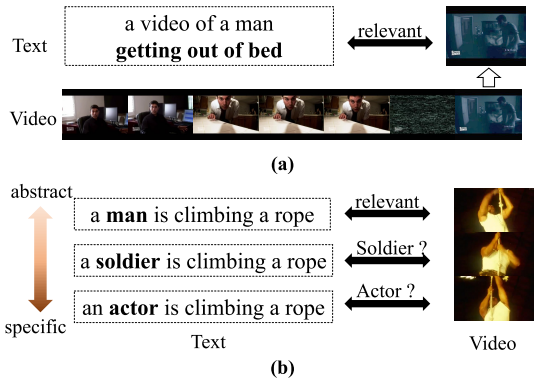

Figure 1: Examples of differences between text and video. (a) The content mentioned in the text appears in a small number of frames in the video. (b) Text contains a different semantic range than video, making it difficult to achieve semantic alignment.

If text and video are divided into multiple parts and then aligned, the problem in Figure 1 will be reduced. Therefore, we can divide video into frames and clips, and text into words and phrases according to the structural hierarchy of them. From a semantic point of view, it also realizes the semantic hierarchy segmentation from abstract to specifics. Moreover, different parts have different extents of importance for retrieval. In Figure 2, we call the parts related to query information *"effective semantics"*. TVR aims to refine and align the effective

*Work was done when Yahan Yu was interning at Minority-Mandarin Translation, Tencent Inc, China.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5547–5557

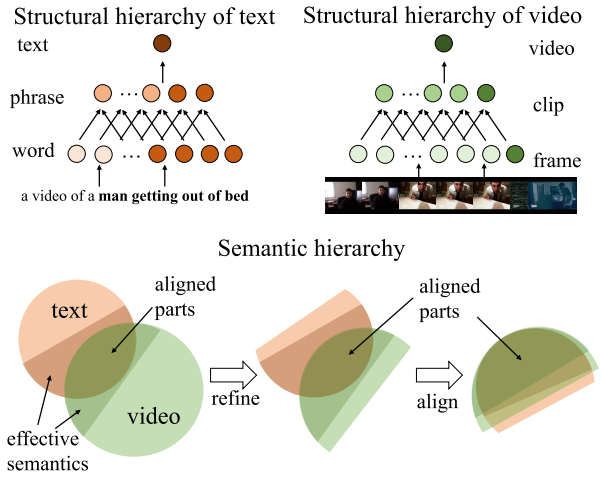

Figure 2: Explanation of structral hierarchy, semantic hierarchy and effective semantics.

semantics of paired text-video through structural hierarchy. Through the step-by-step aggregation of structural and semantic hierarchy, intra-modality feature refining and cross-modality interaction are realized, so that the effective semantics of text and video can be aligned.

To this end, we design a **G**raph-based **H**ierarchical **A**ggregation **N**etwork (GHAN) to refine effective semantics within modality and align effective semantics cross modality. To achieve the former, a token-level weighted network is constructed to aggregate words and frames. As for the latter, we aggregate clips and phrases through graph-based message passing attention network for global-local alignment. Our contributions are summarized below:

- We propose a GHAN method to solve the mismatched problem of information density and capacity between video and text from the perspective of effective information refining and aligning.
- According to the proposed concept of "effective semantics", we design a token-level weighted network to refine intra-modality features, and construct a graph-based message passing attention network for global-local alignment across modality.
- We conducted experiments on the public datasets MSR-VTT-9K, MSR-VTT-7K and MSVD, achieving Recall@1 of 73.0%, 65.6%, and 64.0%, which remarkably boosts the retrieval performance of the current state-of-the-art model CAMoE (Cheng et al., 2021) by 25.7%, 16.5%, and 14.2%.

2 Related Work

Vision-language research. Visual-language research is currently a popular research field, including image-text research (Xing et al., 2021; Frank et al., 2021) and video-text research (Yu et al., 2020). In the early days, visual-language models (Chen et al., 2022) were usually designed independently. Images were usually encoded using handcrafted descriptors (Socher et al., 2013; Elhoseiny et al., 2013). Videos were mostly encoded using 2D/3D spatial-temporal convolution (Tran et al., 2015; Feichtenhofer et al., 2019). Texts were encoded using pre-trained word vectors (Frome et al., 2013) or TF-IDF features (Lei Ba et al., 2015).

Recently, language pretraining models (Devlin et al., 2018) have achieved great success on NLP tasks. Vision-language research has been similarly inspired (Im et al., 2021; Tang et al., 2021; Tan and Bansal, 2020). For image-text pre-training, Lu et al. (2019), Tan and Bansal (2019) used two independent Transformers (Vaswani et al., 2017) to encode image and text respectively. As for Li et al. (2019, 2020); Su et al. (2019), a shared Bert model was then used. CLIP (Radford et al., 2021) learns images directly from text, which leverages a wider range of source of supervision. Lei et al. (2021) proposed an end-to-end method through sparse sampling, which extracted visual and linguistic features with higher efficiency and lower memory usage. Luo et al. (2021) directly used CLIP and trained the model in an end-to-end manner.

Text-Video Retrieval. Although text-image research (Khademi, 2020; Zhang et al., 2020a) has been extensively studied, text-video retrieval is still quite challenging. The earlier works, such as Liu et al. (2019) and Gabeur et al. (2020), solved this task by Mixture-of-Experts (MoE, Ma et al. (2018)), which taked advantage of modalities to integrate generalizable features. Liu et al. (2019) used 7 kinds of feature encoding for video, and then aggregated through Mean Pooling (Lee et al., 2016) or NetVLAD (Arandjelovic et al., 2016), and fused multimodal features together. Gabeur et al. (2020) encoded features through Transformer (Vaswani et al., 2017), calculating the similarity with the text and weighting different modalities according to the text. Based on the idea of NetVLAD (Arandjelovic

et al., 2016), Wang et al. (2021) adaptively aggregated sound, action, scene, speech, OCR, face, etc. in video and a series of shared semantics in text. With the rise of pre-training models, Dzabraev et al. (2021) used the image-text pre-training model CLIP (Radford et al., 2021) to encode the original video, which improved the retrieval accuracy. Luo et al. (2021) transfered CLIP (Radford et al., 2021) to the video domain to solve text-video retrieval task. Cheng et al. (2021) used MoE and CLIP to extract multi-view video representations, including actions, entities, scenes, etc., and then aligned them with corresponding text parts. Our model GHAN also benefits from existing image-text pre-training models, but we are more concerned about the design of interaction between modalities, and extensive ablation studies are conducted to demonstrate the effectiveness of our method.

3 Method

We propose an end-to-end graph-based hierarchical aggregation network, aiming to obtain features with the simplest structure and the most abundant semantics for TVR. Figure 3 describes the architecture of our model named GHAN. Details are described in the rest of this section.

3.1 Encoding Layer

Relevant text and video are fed into our model GHAN in pairs. The pretrained text-image model CLIP (Radford et al., 2021) is effective for the text-video retrieval in this paper. As a result, we utilize CLIP as the backbone to encode the input text and video, which enables us to learn cross-modality interaction with less frames and is more computationally efficient. In our work, we mainly focus on the aggregation and interaction of features rather than the pre-training itself.

For text, Let $T = [w_1, w_2, ..., w_{N_w}]$ be an input text, where $w_i (1 \leq i \leq N_w)$ is the i^{th} word in T. Then we take Bert (Devlin et al., 2018) pretrained by CLIP to encode them as semantic representations:

$$e^w = [e_1^w, e_2^w, ..., e_{N_w}^w] = \text{BERT}(T), \quad (1)$$

where $e^w \in \mathbb{R}^{N_w \times D}$ is the hidden state sequence output by the last layer of Bert, and D means the dimension of each word representation. N_w represents the output sequence length, implying that we encode the original input text into N_w word representations.

Similarly, for video, we define a video as a time sequence of N_f sampled image frames. Let $V = [f_1, f_2, ..., f_{N_f}]$ be a raw video, where $f_j (1 \leq j \leq N_f)$ is the j^{th} frame. We use ViT (Dosovitskiy et al., 2020) pretrained by CLIP as the backbone:

$$e^f = [e_1^f, e_2^f, ..., e_{N_f}^f] = \text{ViT}(V), \quad (2)$$

where $e^f \in \mathbb{R}^{N_f \times D}$ is the sequence of hidden states output by the last layer of ViT. In particular, considering the temporal features between video frames, we add positional encoding to the frame sequence $e^f = e^f + e^{pos}$ to enforce this. N_f denotes the output sequence length, implying that we encode the frame sequence into N_f temporal frame representations.

3.2 Intra-Modality Refining Layer

According to our proposed structural hierarchy, the representations of the lowest hierarchy word and frame have been generated by the Encoding Layer. This level contains the richest original input semantics, so in the process of generating representations at other hierarchies within the modality, we tend to preserve the local effective semantics as much as possible. "Local" can be understood from two perspectives: (i) it contains the original semantic information from the Encoding Layer. (ii) The overall effective semantics consists of multiple discrete local semantics, and the local effective semantics can be regarded as different views of the overall semantics. For example, the reference to the event subject and the reference to the time and place in the text usually belong to different phrase parts. Therefore, in this layer, we aggregate e^w and e^f separately to generate N^p phrase representations and N^c clip representations. In order to achieve the refining of effective semantic information to the next hierarchy, considering that all words and frames do not have the same contribution, we propose a token-level weighted network to perform weighted aggregation of words and frames:

$$e^p = \text{Att}(e^w) m(e^w), \quad (3)$$

$$e^c = \text{Att}(e^f) m(e^f), \quad (4)$$

where $Att(\cdot)$ is the token-level weighted function, $m(\cdot)$ is the stack of linear layers, $e^p \in \mathbb{R}^{N_p \times D}$ is the set of phrase representations, $e^c \in \mathbb{R}^{N_c \times D}$ is the set of clip representations. Although self-attention has been very popular in research work in recent years, in order to keep our model structure

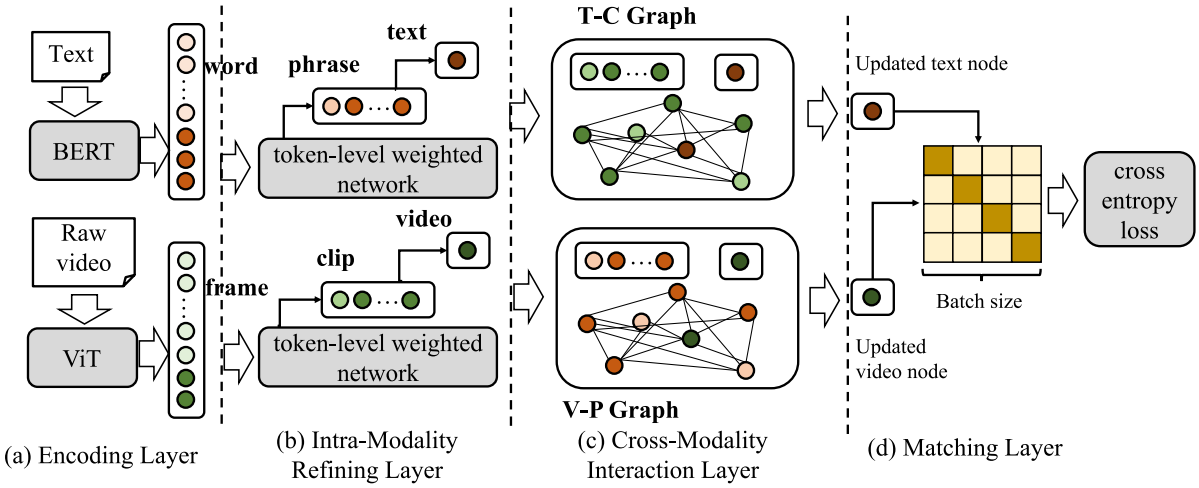

Figure 3: The architecture of our model GHAN. We divide the model into four parts: (a) Encoding Layer: We utilize CLIP as the backbone. (b) Intra-Modality Refining Layer: We preliminarily aggregate words and frames into clips and phrases through a token-level weighted network, refine the effective semantics within modality, and integrate video and text into the same length. (c) Cross-Modality Interaction Layer: we construct a Message Passing Attention network for global-local Alignment (MPAA), which align clips-text and phrases-video independently. The interaction between modalities is performed while retaining the effective semantics within two modalities, and more concise representations are obtained to save unnecessary computation. (d) Matching Layer: Cosine similarity of text and video is calculated to obtain the retrieval result. We use cross-entropy loss to make pairwise matches with greater similarity and others smaller.

from being too complicated, we choose to relatively simplify self-attention. All word and frame weights are generated using only stacking of linear layers, and then aggregated into phrases and clip representations.

$$Att(e^w) = [\sigma(W_{att}e^w + b_{att})]^T, \qquad (5)$$

where σ means softmax activation, W_{att} is the trainable parameter matrix, and b_{att} is the biase. From the aspect of feature dimension, $Att(\cdot) \in \mathbb{R}^{N_p \times N_w}$.Through the visual analysis of the weights after training, it is proved that our network is effective. The aggregation at the structural hierarchy realizes the aggregation at the semantic hierarchy and the refinement of effective semantics.

3.3 Cross-Modality Interaction Layer

So far, we have completed the representation learning of phrases and clips, and the effective information of the original input is weighted and assigned to them. Considering that in this layer, our goal is to achieve cross-modal feature interaction, we first take the average for each phrase and clip to initialize the top-hierarchy text and video representations: $e^t = \frac{1}{N_p}\sum_{i=1}^{N_p} e_i^p$ and $e^v = \frac{1}{N_c}\sum_{j=1}^{N_c} e_j^p$. Where $e^t \in \mathbb{R}^{1\times D}$ is the text initial global representation and $e^v \in \mathbb{R}^{1\times D}$ is the video initial global representation. In terms of information interaction, humans can reason from spatial or semantic dimensions, and graphs can well describe spatial and semantic information(Chen et al., 2018), so that computers can learn to use this information like humans to make inferences. Therefore, we built two global graphs for cross-modal interaction. The construction details are as follows:

(a) T-C graph G_{tc} contains one text node n_0^{tc} and N_c clip nodes n_i^{tc} $(1 \le i \le N_c)$. Specially, $e^t \in \mathbb{R}^{1\times D}$ denotes the representation of n_0^{tc}, $e_i^c \in \mathbb{R}^{1\times D}(1 \le i \le N_c)$ denotes representation of the i^{th} clip node n_i^{tc}. We add edges between any two nodes in the graph. T-C graph aligns text global semantics using clip local semantics.

(b) V-P graph G_{vp} contains one video node n_0^{vp} and N_p phrase nodes $n_j^{vp}(1 \le j \le N_p)$. Specially, $e^v \in \mathbb{R}^{1\times D}$ denotes the representation of n_0^{vp}, $e_j^p \in \mathbb{R}^{1\times D}(1 \le j \le N_p)$ denotes representation of the j^{th} phrase node n_j^{vp}. We add edges between any two nodes in the graph. V-P graph aligns video global semantics using phrase semantics.

By constructing two independent graph, the global information of one modality is aggregated

with the local information of the other modality. At the same time of interaction at the structural hierarchy, the effective semantic alignment between modalities is realized. In the aggregation of graph nodes we treat all node and edge types as the same, i.e. we build homogeneous graphs. We also explore other ways of constructing graph, which are described in Section 4.

The Message Passing Neural Network (MPNN, Zhang et al. (2020b)) is a framework based on the core idea of recursive neighborhood aggregation, where nodes can pass messages to each other, and the representation of each node is updated according to the messages received from its neighbors. MPNN consists of two phases: aggregation phase and the readout phase. The aggregation phase including Aggregation Function and Combination Function runs for t time steps in total. The Aggregation Function aggregates features of neighbor nodes, ready to be passed to the central node. The Combination Function updates the node representation, combining the representation of the node with the message obtained from the Aggregation Function. The readout phase obtains graph-level representations through the Readout Function for subsequent classification or regression tasks. Considering the graph attention mechanism proposed by Graph Attention Networks (GAT, Veličković et al. (2017)), we design a Message Passing Attention network for global-local Alignment (MPAA) as shown in Figure 4. We apply the same MPAA to G_{tc} and G_{vp}. For a graph G, we design the following Aggregation Function:

$$m_{k_i}^t = \sum_{j=0}^{|\mathcal{N}_i|} \alpha_{k_{ij}} h_j^{t-1} W_a, \tag{6}$$

where h_j^{t-1} means the hidden representation of j_{th} node at the $t-1$ time step, W_a is the trainable parameter matrix in aggregation function, $\alpha_{k_{ij}}$ means the k_{th} attention coefficient between i_{th} node and j_{th} node, k means a total of k attention mechanisms need to be considered, $\mathcal{N}_i$ means the set of neighbor nodes of the i_{th} node, $\mid \mathcal{N}_i \mid$ is the number of nodes in $\mathcal{N}_i$, and $m_{k_i}^t$ is the message vector. More specifically, the k_{th} attention coefficient $\alpha_{k_{ij}}$ expands as follows:

$$\alpha_{k_{ij}} = \frac{\exp\left(\delta\left(a_k^T\left[h_i^{t-1}W_a \| h_j^{t-1}W_a\right]\right)\right)}{\sum_{v=0}^{|\mathcal{N}_i|}\exp\left(\delta\left(a_k^T\left[h_i^{t-1}W_a \| h_v^{t-1}W_a\right]\right)\right)}, \tag{7}$$

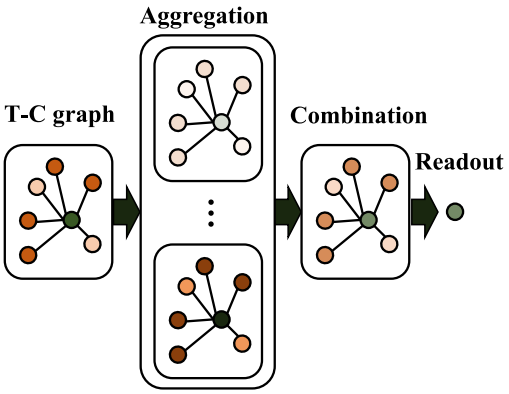

Figure 4: Process flow of Message Passing Attention network for global-local Alignment (MPAA). Taking text as an example, clip nodes processed by multiple attention mechanisms is transmitted in the aggregation and connected. Then a single attention is used to update the text representation in the readout phase.

where h_j^{t-1} is the j_{th} neighbor node's hidden representation of the i_{th} node, a_k is the trainable parameter matrix, and δ is LeakyReLu activation. We design the Combination Function as follows which takes $m_{k_i}^t$ as input:

$$h_i^t = \|_k[\phi(m_{k_i}^t + h_i^{t-1}W_c)], \tag{8}$$

where ϕ is ELU activation, W_c is the trainable parameter matrix in combination function, $\|_k$ means concatenating the outputs of the k attention mechanisms, and h_i^t means the updated representation of i_{th} node at the t time step. The purpose of our setting W_c is to make the updated representations not deviate too much from the original representations and enhance the robustness of the MPAA. Finally, for the Readout Function, we design as follows:

$$r = \sigma[\phi(\sum_{p=0}^{|\mathcal{N}_0|} \alpha_r h_p^t W_{r_1} + h_0^t W_{r_2})], \tag{9}$$

where $h_0^t \in \mathbb{R}^{1\times D}$ means the hidden representation of the text node n_0^{tc} or the video node n_0^{vp}, $\mathcal{N}_0$ means the set of neighbor nodes of n_0^{tc} or n_0^{vp}, $\mid \mathcal{N}_0 \mid$ is the number of nodes in $\mathcal{N}_0$, h_p^t is the representation of nodes in $\mathcal{N}_0$, α_r means the attention coefficient with single head, ϕ means ELU activation, σ means softmax activation, and W_{r_1} and W_{r_2} are the trainable parameter matrixs in readout function. Finally, we set $e^t = r$ in graph G_{tc}, and $e^v = r$ in graph G_{vp}.

3.4 Matching Layer

In this layer, we define two types of retrieval tasks: retrieving video with text (T2V) and retrieving text

Method	T2V				V2T			
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R
CE (Liu et al., 2019)	20.9	48.8	62.4	6.0	20.6	50.3	64.0	5.3
MMT (Gabeur et al., 2020)	26.6	57.1	69.6	4.0	27.0	57.5	69.7	3.7
Frozen (Tsimpoukelli et al., 2021)	31.0	59.5	70.5	3.0	-	-	-	-
MDMMT (Dzabraev et al., 2021)	38.9	69.0	79.7	2.0	-	-	-	-
CLIP4Clip (Luo et al., 2021)	44.5	71.4	81.6	2.0	42.7	70.9	80.6	2.0
CAMoE (Cheng et al., 2021)	47.3	74.2	84.5	2.0	49.1	74.3	84.3	2.0
GHAN(ours)	**73.0**‡	**99.7**‡	**99.9**‡	**1.0**	**74.1**‡	**99.2**‡	**99.9**‡	**1.0**

Table 1: Comparison of different methods on MSR-VTT-9K. We perform a significance test which show the improvements over baseline and SOTA are both statistically significant ("‡" indicates $p < 0.01$).

with video (V2T). The goal of the retrieval task is to interact with representations of text and video so that the larger the pairwise similarity, the smaller the others. So we apply cross-entropy loss (Zhai and Wu, 2018) to this task. There are N text-video pairs in a batch B that treat text-video pairs as positive samples and others as negative samples, and define the overall loss as the average of the two retrieval tasks:

$$Loss_{t2v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(d(e_i^t, e_i^v))}{\sum_{j=1}^{B}\exp(d(e_i^t, e_j^v))}, \tag{10}$$

$$Loss_{v2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(d(e_i^t, e_i^v))}{\sum_{j=1}^{B}\exp(d(e_j^t, e_i^v))}, \tag{11}$$

$$Loss_{total} = (Loss_{t2v} + Loss_{v2t})/2, \tag{12}$$

where $d(\cdot)$ is the cosine similarity function used for text and video distance measurement. The cross-entropy loss enables our model to learn matching the most relevant text and video. The loss function uses a symmetric cross-entropy loss over similarity scores. Every text and video are calculated similarity to all videos or texts, which should be maximum in ground truth pairs. When the cosine similarity between embeddings output by the model is the largest, the loss is the smallest. This can meet the needs of model training.

4 Experiment

4.1 Datasets

MSR-VTT-9K (Gabeur et al., 2020). MSR-VTT (Xu et al., 2016) consists of 10k videos ranging in length from 10 to 30 seconds, each paired with approximately 20 texts. In MSR-VTT-9K we use the training split in Gabeur et al. (2020) which consists of about 9k videos and 180k texts and the 1K-A split test set (Yu et al., 2018) which contains 1k selected text-video pairs.

MSR-VTT-7K (Miech et al., 2019). We use a training split of Miech et al. (2019) which contains approximately 7k video sets and 140k texts and split the entire dataset into 7k for training and 3k for testing.

MSVD (Chen and Dolan, 2011) contains 1,970 videos ranging in length from 1 second to 60 seconds and about 120k texts. Each text describes a video. The training, validation, and test datasets consist of 1,200, 100, and 670 videos.

4.2 Experimental Settings

In the Encoding Layer, we process raw videos and texts using the same ViT (ViT-B/32) and Bert in CLIP, and initialize all encoder parameters from CLIP's pretrained weights. We sample 1 frame per second and resize each frame to 224×224. Both frame and word embeddings have dimension $D = 512$ and use the same logit scaling parameter as CLIP.

In the Intra-Modality Refining Layer, $Att(\cdot)$ uses a linear layer with the input dimension 512 and output dimension N_p for words and output dimension N_c for frames. $m(\cdot)$ uses linear layers with input dimension of 512, hidden size of 1024 and output dimension of 512.

In the Cross-Modality Interaction Layer, there are $(N_c + 1)$ nodes n_i^{tc} and $(N_c + 1)^2$ edges in T-C graph and $(N_p + 1)$ nodes n_j^{vp} and $(N_p + 1)^2$ edges in V-P graph. The two graphs use two same MPAAs. Their input and output dimensions are 512, hidden size is 256, dropout is 0.2, attention heads are 8 and Xavier initialization method is used.

In the Matching Layer, we follow the evaluation metric (Luo et al., 2021) and report recall with rank K (R@K),and median rank (Med R). Higher R@K and lower Med R indicate better performance.We set batch size to 128 for all experiments, $N_w = N_v = 32$, $N_c = N_p = 6$, learning rate to $1e-7$ for the CLIP initialization weights and $1e-4$ for

Method	T2V				V2T			
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R
UniVL (Luo et al., 2020)	21.2	49.6	63.1	6.0	-	-	-	-
ClipBERT (Lei et al., 2021)	22.0	46.8	59.9	6.0	-	-	-	-
MDMMT (Dzabraev et al., 2021)	26.6	57.1	69.9	4.0	-	-	-	-
Support (Patrick et al., 2020)	27.4	56.3	67.7	3.0	26.6	55.1	67.5	3.0
Frozen (Tsimpoukelli et al., 2021)	31.0	59.5	70.5	3.0	-	-	-	-
CLIP4Clip (Luo et al., 2021)	42.1	71.9	81.4	2.0	-	-	-	-
CAMoE (Cheng et al., 2021)	48.8	75.6	-	-	50.3	74.6	-	-
GHAN(ours)	**65.6**‡	**95.8**‡	**98.8**‡	**1.0**	**65.7**‡	**96.4**‡	**99.0**‡	**1.0**

Table 2: Comparison of different methods on MSR-VTT-7K. We perform a significance test which show the improvements over baseline and SOTA are both statistically significant ("‡" indicates $p < 0.01$).

Method	T2V				V2T			
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R
CE (Liu et al., 2019)	19.8	49.0	63.8	6.0	-	-	-	-
Support (Patrick et al., 2020)	28.4	60.0	72.9	4.0	-	-	-	-
Frozen (Tsimpoukelli et al., 2021)	33.7	64.7	76.3	3.0	-	-	-	-
CLIP4Clip (Luo et al., 2021)	46.2	76.1	84.6	2.0	56.6	79.7	84.3	1.0
CAMoE (Cheng et al., 2021)	49.8	79.2	87.0	-	-	-	-	-
GHAN(ours)	**64.0**‡	**99.0**‡	**99.0**‡	**1.0**	**67.0**‡	**99.7**‡	**99.8**‡	**1.0**

Table 3: Comparison of different methods on MSVD. We perform a significance test which show the improvements over baseline and SOTA are both statistically significant ("‡" indicates $p < 0.01$).

others. We optimized our model for 5 epochs using the Adam optimizer. We use 4 A100 GPUs for training, and the training duration is about $4h$.

4.3 Main Results

Table 1, 2 and 3 list the results of the comparative models and our model GHAN on MSR-VTT-9K, MSR-VTT-7K and MSVD, respectively. We have the following observations:

(1) The models of hierarchical aggregation (Cheng et al., 2021; Dzabraev et al., 2021) generally perform better. The Mixtures of Experts method that integrates features from various modalities is also a hierarchical aggregation architecture, which makes the semantic interaction across modalities more sufficient through hierarchical aggregation of features.

(2) By comparing our model with Dzabraev et al. (2021); Cheng et al. (2021), we build a graph structure in hierarchical aggregation networks for cross-modal information aggregation, and experiments show that the effect is remarkable.

(3) The image-text pre-trained model is helpful for improving the results. Compared with Dzabraev et al. (2021) and Gabeur et al. (2020); Liu et al. (2019), fusing pre-trained model has advantages.

(4) GHAN achieves the best results among all models. We attribute this to the design of the hierarchical structure, the cross-modal interaction of the graph structure, the intra-modal feature refining, and the transfer of the image-text pre-trained model. Further studies are shown below.

4.4 Effects of each layer of GHAN

The baseline we use simply average the obtained frame and word features from the output of the pre-trained model, and then calculate the cosine similarity for retrieval. As can be seen from the Table 4, our subsequent measures of intra-modal aggregation(token-level weighted network) and cross-modal interaction (Message Passing Attention Network for Global-Local Alignment, MPAA) are effective. w/o inter refers to only using the encoding layer and token-level weighted network for training, and w/o intra refers to using the encoding layer and MPAA for training. Neither single intra-modal aggregation nor single cross-modal aggregation can be used to achieve good results. It demonstrates that simple linear layer stacking fails to adequately interact with semantics across modalities. Meanwhile, direct cross-modal aggregation is easy to lose fine-grained features, and effective semantics are disturbed by a large number of irrelevant semantics. Thus, the effects of each layer design of our model is verified.

model	T2V			
	R@1	R@5	R@10	Med R
baseline	37.0	64.2	74.7	3.0
w/o inter	44.0	71.2	81.3	2.0
w/o intra	43.2	72.0	81.5	2.0

Table 4: Ablation studies of interaction layers in GHAN

model	T2V			
	R@1	R@5	R@10	Med R
Ho-1Graph	49.1	89.1	97.6	2.0
Ho-2Graph	73.0	99.7	99.9	1.0
He-1Graph	49.5	88.0	97.5	2.0
He-2Graph	73.6	99.0	99.9	1.0

Table 5: Effects of different graph structures

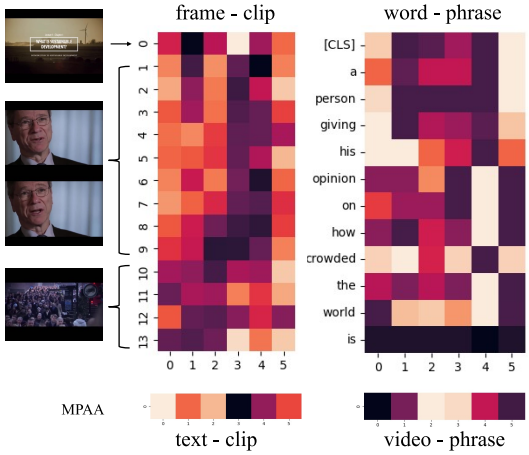

Figure 5: Case analysis. Light colors indicate high weights.

4.5 Effects of Different Graph Structures

In the Table 5, we explore four graph structures for updating text and video representations, namely Ho-1Graph, Ho-2Graph, He-1Graph and He-2Graph. Ho-1Graph refers to the construct of one homogeneous T-V-C-P graph for all four kinds of nodes, and Ho-2Graph refers to the method finally selected in this paper, which constructs two graphs T-C graph and V-P graph respectively. He-1Graph & He-2Graph retain the corresponding node connections, and treat nodes and edges at different hierarchies and modalities as a difference. Firstly, the influence of homogeneous graph and heterogeneous graph is analyzed. The homogeneous graph still uses MPAA to achieve message-passing. As there are various types of nodes and edges in the heterogeneous graph, we choose R-GCN (Schlichtkrull et al., 2018) for training. From the Table 5, it can be concluded that the method of heterogeneous graph is more effective, but training time in 5 epochs of heterogeneous graph is 9.2h, while the homogeneous graph only needs 4h, so we finally ignored the slight improvement brought by heterogeneous graph and chose to build the same composition.

After choosing the way of homogeneous graphs, we analyze whether to build features into one graph or two graphs. It is better to aggregate text and video separately according to the Table 5. As the global feature already contains the local features of the same modality, the structure of two graphs can make the effective semantics of the global feature directly align with the local effective semantics of another modality, bringing about a significant improvement in the results.

4.6 Analysis

In this section, we conduct further analyses to demonstrate the inner workings of GHAN. We select a text-video pair to visualize the weights of the aggregation process, as shown in Figure 5.

Intra-Modality Refining We first explore the capability of the proposed method to refine semantics in the structural hierarchy. Frames contains 3 scenes: opening subtitles, person speaking and crowd. Frame-clip weight shows that these scenes are well refined and aggregated into different clips respectively. Word-phrase weight shows the information in the 6 phrases: *a person giving his crowded/ his crowded world/ his opinion world/ crowded world/ opinion on how world/ a person giving his crowded*, effectively dividing the semantics.

Cross-Modality Interaction We provide a further investigation on whether the designed MPAA has aligned clips and phrases containing different semantics across modalities. We visualize one head of the multi-head attention in MPAA. As shown in Figure 5, Text-clip weight shows that the fourth clip is filtered during the alignment process, whose main semantics is the scene of the opening subtitles and is irrelevant to the text. Video-phrase weight shows that video pays more attention to phrases which express *his opinion world/ crowded world/ opinion on how world*. This demonstrates that MPAA can satisfactorily achieve cross-Modality alignment and interaction, leading to the inherent superiority in text-video retrieval.

5 Conclusion

In this paper, we propose a graph-based hierarchical aggregation network named GHAN for text-video retrieval. We noticed that text and video have structural and semantic hierarchies. In this regard, we propose the concept of effective semantics, which maps retrieval tasks to refine and align effective semantics between modalities. We design a token-level weighted network to refine intra-modality features, and build a message passing attention network for global-local alignment across modality. Our model significantly improves on multiple datasets. More experiments indicate that our model can well achieve semantic refining and alignment.

6 Limitations

Our model simply performs text-video retrieval and does not process speech information. Speech information in the video, such as character dialogue or narration, can enrich the semantics of the video. If the speech information is added, it is envisaged that the retrieval accuracy can be improved, and more practical tasks can be designed, such as mutual retrieval of text, voice, video, and images. On the other hand, our approach lacks alignment between texts and semantics within frames. Some objects may occupy a small proportion of a frame, resulting in retrieval failure. In future work, we plan to take advantage of more modal informationbuild a unified model that simultaneously implements retrieval tasks and generation tasks.

References

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2022. VLP: A survey on vision-language pre-training. *CoRR*, abs/2202.09061.

Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. 2018. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7239–7248.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363.

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403.

Mahmoud Khademi. 2020. Multimodal neural graph memory networks for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7177–7188.

Chen-Yu Lee, Patrick W Gallagher, and Zhuowen Tu. 2016. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In *Artificial intelligence and statistics*, pages 464–472. PMLR.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. 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*, pages 7331–7341.

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 4247–4255.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11915–11925.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 525–535.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. 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.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487.

Andrew Zhai and Hao-Yu Wu. 2018. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*.

Bowen Zhang, Hexiang Hu, Vihan Jain, Eugene Ie, and Fei Sha. 2020a. Learning to represent image and text with denotation graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 823–839.

Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020b. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339.