Interpretable Video-Text Alignment (VTA) for Cross-Modal Retrieval

Tsung-Shan Yang* Yun-Cheng Wang* Chengwei Wei* Suya You† C.-C. Jay Kuo*
* University of Southern California, USA
{tsungsha, yunchenw, chengwei, jckuo}@usc.edu
† DEVCOM Army Research Laboratory, USA
suya.you.civ@army.mil

***Abstract*—Aligning representations between the video and text domains is crucial in video-text retrieval. Most deep learning algorithms conceal this mechanism in a latent space through end-to-end optimization. In addition, the computational requirements for training and inference increase exponentially with the video length. To address interoperability and efficiency, we propose an interpretable video-text alignment (VTA) method in this work. Through keyframe selection and genre clustering, we reduce the number of trainable parameters to just 3% compared to CLIP [1] backbone training while maintaining constant time and space complexity during inference.**

Furthermore, our alignment method connects embedding spaces of different modalities. During the alignment process, the conditional probabilities for a given object or phrase are estimated and aggregated, providing probabilistic support for our predictions. Hence, the alignment method is fully transparent. Experimental results on the MSR-VTT dataset demonstrate that the proposed VTA method offers state-of-the-art performance. Moreover, our alignment method does not demand fine-tuning of the pre-trained encoder, thereby avoiding bias from small paired training datasets.

***Index Terms*—Video-Text Retrieval, Video Content Understanding, Multimodal Alignment.**

I. Introduction

In today's fast-paced digital content creation and consumption environment, efficiently retrieving relevant information from various multimedia sources is more important than ever. An example is the video-text retrieval (VTT) task, which finds the most relevant video related to a given sentence and vice versa. It can summarize video content in just a few words or find a highly relevant video clip based on a sentence description. The MSR-VTT dataset [2] is illustrated in Figure 1, where a video clip has five or more captions. A VTT algorithm must retrieve the most relevant counterpart from the other modality when provided with either a clip or a caption.

Although Large Multimodal Models (LMMs) exhibit remarkable performance in visual content understanding, they face three challenges: (1) data scarcity, (2) high training and inference costs, and (3) a black-box process. To address these issues, we propose an efficient and transparent video-text alignment (VTA) method, which bridges the embedding spaces in the video and text domains.

The proposed VTA method has three advantages.

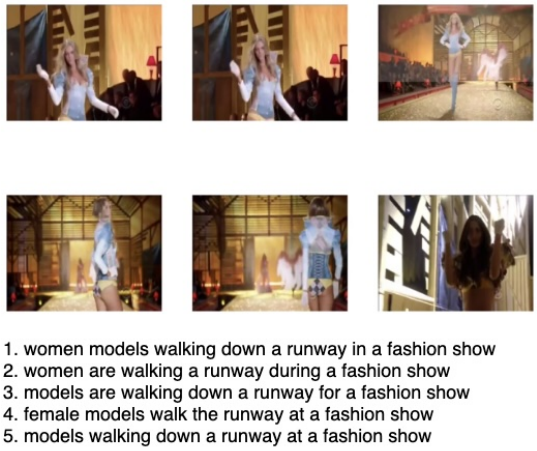

Fig. 1. A video-text pair in the MSR-VTT dataset [2], where each video clip has five or more captions.

- A limited number of paired video-text datasets are available compared to the larger volume of paired image-text datasets. If a model is trained on a limited dataset, it faces a high risk of bias. To mitigate this risk, VTA keeps the visual and textual encoders frozen during training, preventing bias from affecting the model.
- The computational cost of processing videos increases exponentially compared to that of images. In addition to the spatial components within individual frames, videos also contain temporal information. As the number of frames increases, algorithms need to perform more operations to extract features, which can significantly increase computational demands. To address this challenge, VTA employs a keyframe selection module that minimizes repetitive computations by inferring the model based on a specific number of keyframes.
- Black-box LMMs can cause unexpected errors. To allow a transparent pipeline, VTA decomposes the video-text retrieval process into a series of probabilistic predictions. All intermediate results are meaningful. Specifically, VTA utilizes an object detector and part-of-speech (POS) tagging to conceptualize videos and sentences in the joint domain. We associate similar concepts with the information in the visual and textual domains.

The VTA method consists of two stages: 1) removal of irrelevant samples and 2) ranking relevant samples. They will be elaborated on in Section III. Our main contributions include the following.

- VTA reduces the number of trainable parameters to 3%, compared to fine-tuning the visual and textual encoders.
- VTA freezes the encoders in the training stage, which helps to maintain generalizability across domains.
- VTA illustrates the retrieval process with sequential intermediate results. By assembling the results, retrieval can be viewed as a concept-matching process.

II. Related Work

Video-text retrieval relies on encoders that process visual and textual data. The researchers initially developed jointly trained encoders using paired image and text datasets. This trend has since expanded from single images to sequential frames. Various time sequence structures, such as recurrent neural networks (RNNs) and attention mechanisms, have been proposed to manage temporal information effectively. With the recent advancement of Large Language Models (LLMs), research has increasingly focused on aligning visual features with language tokens. However, the reasoning process remains obscured in a latent space. Kuo et al. [3] proposed Green Learning, which avoids backward propagation and provides understandable intermediate results to demystify the decision-making process.

A. Multimodal Models

Torabi et al. [4] and Yu et al. [5] discuss the joint representations across video and text domains. The foundation of today's dual-encoder architectures is built on a combination of convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks. Metric learning plays a crucial role in creating a shared latent space. Hadsell et al. [6] propose the idea of contrastive learning. The loss formulation aims to reduce the distance in the latent space for similar samples and to increase the distance for different samples.

Implementing the training scheme on a large-scale dataset, Radford et al. [1] show the effectiveness of large pre-trained models in enhancing joint representations. Portillo et al. [7] demonstrate a straightforward frame-wise inference approach for video-to-text retrieval, demonstrating the generalizability of the CLIP model.

B. Visual Language Models

Large language models (LLMs) achieve impressive reasoning and understanding in natural language processing. Research focuses on leveraging the power of large foundation models for image and video understanding applications. Ma et al. [8] introduce the dual-encoder structure to align the text and video representation. However, the jointly training process relies on the dense attention mechanism on the frame-wise patches and word tokens. The dense inference structure may involve repetitive patch computation between frames.

Instead of aligning two encoders, a pipeline has been developed that unifies the representations. For example, Wang et al. [9] built unified representations in visual and textual domains. The decoders adopted token embedding from both domains, and another model mapped visual and textural tokens onto a unified decoding space.

C. Green Learning

To address the lack of transparency and efficiency in deep learning methods, Kuo et al. [3] proposed a new learning paradigm called Green Learning. It adopts a modularized and feed-forward training process. Users can understand all intermediate results. Green learning consists of: 1) representation learning, 2) feature learning, and 3) decision learning. For (1), multistage PCA-based transforms [10] enable multiple joint spatial-spectral representations. For (2), statistics-based feature ranking tools are performed via the discriminant feature test or the relevant feature test [11]. For (3), the XG classifier or regressor is commonly used.

Recent research has paid attention to the two modalities of transparent interactions of image and text. For example, Yang et al. [12] detected human-object interaction (HOI) types through grouping and encoding. The new model significantly reduces computation complexity while offering competitive performance. Furthermore, Yang et al. [13] proposed the stage-wise alignment for image-text retrieval. In this work, we generalize the problem by extending the visual modality from images to videos.

III. Proposed VTA Method

In the proposed learning paradigm, we divide the retrieval task into two stages: 1) removal of irrelevant samples and 2) ranking of relevant samples. The first stage explores similarities in visual and textual pairs to remove unlikely candidates in the target domain. Then, we can train a smaller model to rank the remaining samples in the second stage. As shown in Figure 2, the removal process compares the key concepts in the videos and captions, respectively. We perform keyframe selection for the former to extract representative visual information before object detection. For the latter, we use the Part-of-Speech (POS) tagging to analyze keywords in captions. By comparing the co-occurrence of detected objects and keywords, we exclude irrelevant pairs as presented in Sec III-A. The ranking process sorts candidates according to the clustering of genres and trains projection matrices with paired data of similar genres based on the embeddings of frozen encoders. Although a simple projection matrix may struggle with complex retrieval tasks, it performs well against homogeneous subsets. More details will be given in Sec . III-B.

A. Stage 1: Irrelevant Samples Removal

We analyze the co-occurrence of objects in video clips and keywords in captions. Thus, we can eliminate unlikely pairs of clips and captions by filtering out object-keyword pairs of low occurrence. The occurrence is proportional to the conditional probability between objects and keywords. The

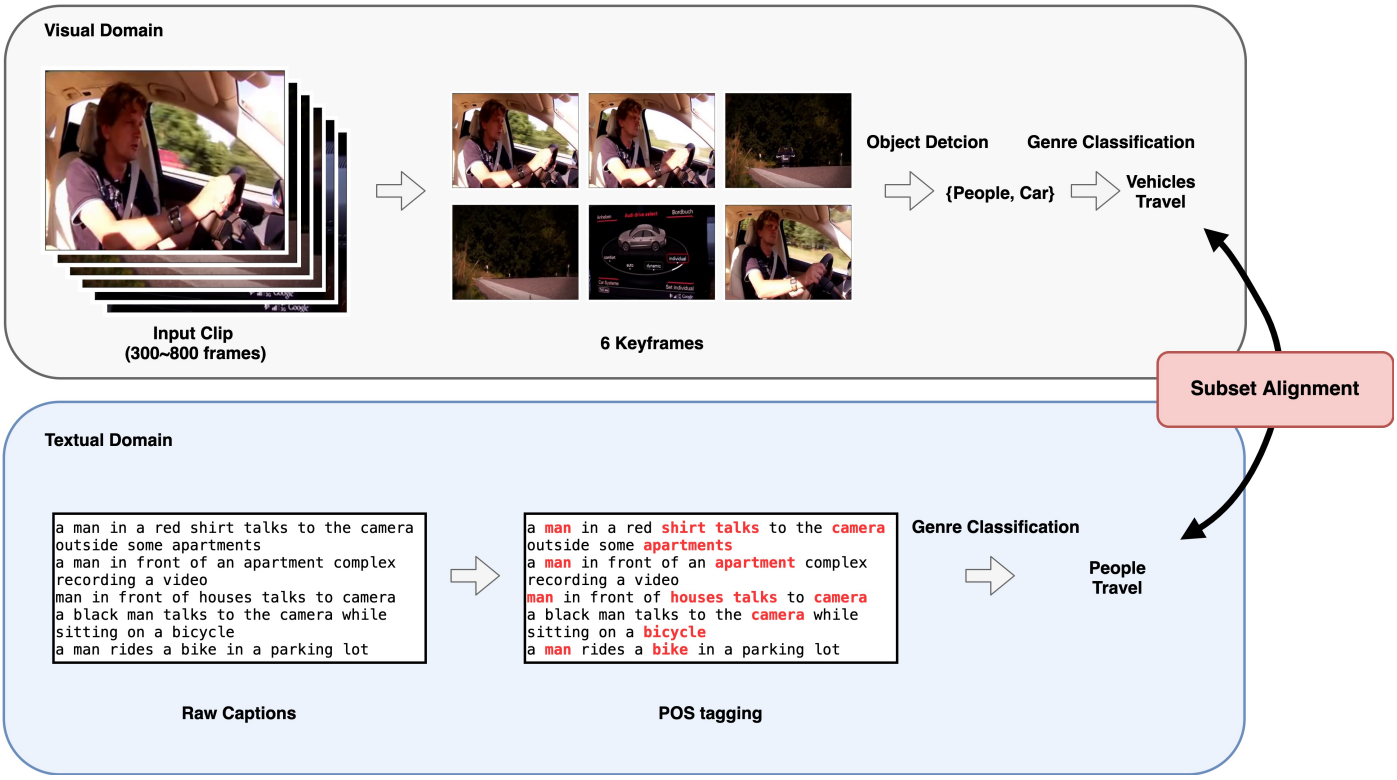

Fig. 2. The proposed VTA process involves several steps. It uses keyframe selection and object detection for genre classification in the visual domain, and identifies keywords through part-of-speech (POS) tagging in the textual domain. Then, it trains an alignment module in each genre group.

process of identifying a set of detected objects from video clips and building noun and verb sets from captions is detailed in Sections III-A1 and III-A2, respectively.

1) Keyframe Selection and Object Detection from Videos: As shown in Figure 3, the visual concept extraction module consists of 1) keyframe selection and 2) object detection. The length of the input clips ranges from 10 to 30 seconds. Each clip contains 200 to 800 frames at 24 frames per second. Applying a pre-trained object detector to each frame is computationally intensive. To reduce complexity, we use the histogram difference to find shot changes in a clip and select the frame in the middle of each shot as a keyframe. Then, the set of keyframes represents the query clip. The histogram difference can be computed as

$$\chi(\text{frame}_i, \text{frame}_j) = \sum_{b \in \{bins\}} \frac{(\text{hist}_{i,b} - \text{hist}_{j,b})^2}{\frac{1}{2}(\text{hist}_{i,b} + \text{hist}_{j,b})}, \quad (1)$$

where frame_i and frame_j are two consecutive frames in the clip, and $\text{hist}_{i,b}$ is the number of pixels in a grayscale partitioned bin, b, of frame_i. We set the bin number to eight. The histogram difference is the sum of the differences in all bins.

Frames with the five most significant histogram differences are identified as shot changes. Five-shot transits and the starting and ending time stamps can segment the original clip into six non-overlapping intervals. The keyframes are the middle frames in the intervals, reducing the number of pre-trained object detection inferences from hundreds to six.

2) Keywords Selection from Captions: To establish relations between detected objects and captions, we utilize Part-of-Speech (POS) tagging to classify words in captions and select the 100 most frequent verbs and nouns to build the concept set. Next, we use the video and caption pairs to construct co-occurrence matrices that indicate the joint probabilities between detected objects and keywords in captions.

For a given video, we can formulate the matching probability of a keyword as

$$\begin{aligned} & P(\text{caption}|\text{video}) \\ & \sim P(\{\text{nouns}\}, \{\text{verbs}\}|\{\text{detections}\}) \\ & = \prod_{\substack{n \in \{\text{nouns}\} \\ obj \in \{\text{detections}\}}} P(n|obj) \prod_{\substack{v \in \{\text{verbs}\} \\ obj \in \{\text{detections}\}}} P(v|obj), \end{aligned} \quad (2)$$

where $\{\text{nouns}\}, \{\text{verbs}\}$ denote the word sets with POS tagging labels 'noun' and 'verb' in a caption, $\{\text{detection}\}$ is the detection result from its keyframes, and n, v, obj stands for a specific concept. The normalized values in the co-occurrence matrix approximate the probabilities. In this stage, we approximate the joint probability with the product of probabilities under the independent assumption. Although this is a rough approximation, our objective here is to prune the candidate set by removing unlikely captions to have a smaller training subset for sample ranking in the second stage for higher efficiency.

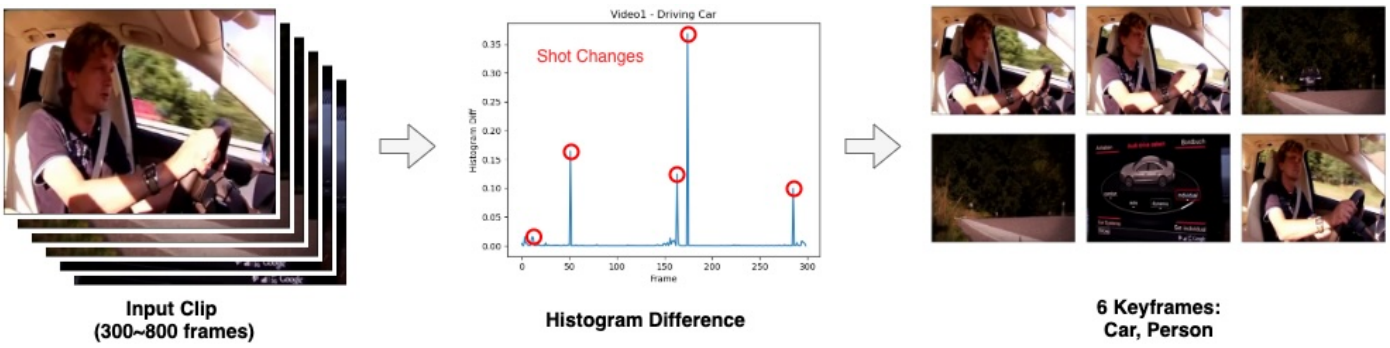

Fig. 3. The pipeline of keyframe selection and object detection from keyframes.

B. Stage 2: Relevant Samples Ranking

Large multimodal models (LMMs) perform remarkably in various applications. However, their high computational complexity and lack of interpretability pose challenges. To build a transparent and efficient solution, we classify clips and captions into subgroups by genre. It allows smaller modules to be trained in each subgroup, reducing computational burdens. The genre indicates the semantic categories to be retrieved when given keywords and objects. The genre classification and contrastive learning building blocks are discussed in Sections III-B1 and III-B2, respectively.

1) Genre Classification: Instead of building a black-box model to manage all types of video clips, we cluster a diverse dataset into homogeneous genre types. The genre labels indicate the characteristics of the clip's content. The video clips in the MSR-VTT [2] dataset have 20 genre labels. Since genre labels are not mutually exclusive, we cluster them into several subsets and consider the top-2 subsets in inference.

We use features from the latent space of the DETR model [14] and detection results from the previous removal stage to train an XGBoost classifier [15]. As illustrated in Figure 4, the genre labels reveal confusion pairs. For example, 'food' and 'cooking' are semantically similar, as shown in the confusion matrix. To address this, we merge some confusion pairs to ensure that candidates are not excluded from the classification results.

To group genre labels, we represent labels as vertices in a graph. The number of confusion pairs serves as the edge weight between two labels. If two labels have fewer confusing pairs, their edge can be pruned to produce a sparser graph. We can run a graph partitioning algorithm to divide the whole graph into several disconnected subgraphs, leading to fewer non-overlapping genre types. We can group captions in the textual domain using the same genre types.

2) Contrastive Learning: Reducing the number of clip/caption pairs in each genre type will make it more convenient to align the visual and textual data in each genre type. We rank the similarity of the cross-domain data by contrastive learning [16]. The loss function is written as

$$\mathcal{L}_{con} = -\log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_k)/\tau)}, \quad (3)$$

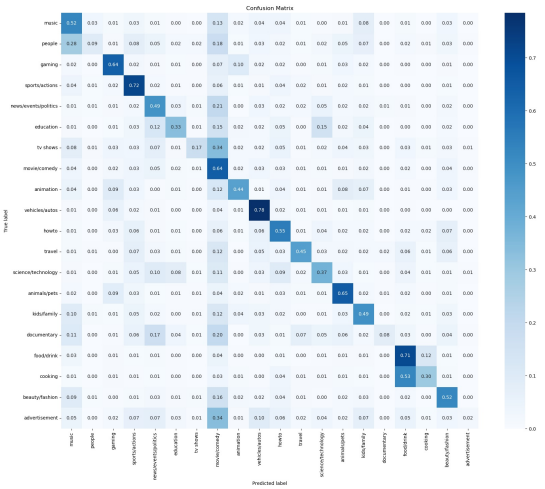

Fig. 4. The confusion matrix among 20 genre labels.

where $z \in \mathbf{R}^{d_{joint}}$ is a feature vector of dimension d_{joint}, (i, j) is a paired image and sentence in the sampled batch, N is the batch size, $\tau \in \mathbf{R}$ is the temperature hyperparameter, and $\mathbb{1} \in \{0, 1\}$ is an indicator function, whose value is one if $[k \neq i]$. The objective function maximizes the similarity of relevant image-text pairs and prevents negative image-text pairs from being close in the latent space. We adopt cosine similarity as the similarity metric, namely $sim(u, v) = \frac{u \cdot v}{\|u\| \|v\|}$.

Aiming for efficiency, we do not fine-tune the pre-trained encoders. Instead, we train a set of lightweight linear transformations that project visual and textual features onto a joint embedding space as:

$$\begin{aligned} z_{vis} &= A \times \mathcal{F}(\text{Frame}) \in \mathbf{R}^{\mathbf{d_{joint}}} \\ z_{txt} &= B \times \mathcal{G}(\text{Caption}) \in \mathbf{R}^{\mathbf{d_{joint}}}, \end{aligned} \quad (4)$$

where z_{vis} and z_{txt} are the image and text embeddings in the joint space, $\mathcal{F}$ and $\mathcal{G}$ are the frozen image and text encoder models, and $A \in \mathbf{R}^{\mathbf{d_{joint}} \times \mathbf{d_{vis}}}$ and $B \in \mathbf{R}^{\mathbf{d_{joint}} \times \mathbf{d_{text}}}$ are trainable matrices that match the output dimensions of the frozen encoders.

Building a video encoder that supports multimodal applica-

tions is computationally intensive. We developed a keyframe selection procedure to reduce training costs that summarizes representative information from a clip. It avoids constructing a video encoder from scratch while maintaining essential information across multiple frames. The alignment is based on the image-level encoder, preventing an explicit use of temporal information. We can concentrate on more relevant pairs for contrastive learning by eliminating irrelevant candidates earlier.

Furthermore, we adopt two training strategies, weighted hard negative mining and self-adaptive temperature, to improve the convergence of the projection matrices. We assign a weight to the top 30% negative pairs for the former and implement a dynamic temperature control for the latter, where the temperature value is determined by the mean of the training set from the previous iteration, in training.

IV. Experiments

A. Experiment Settings

We perform video-to-text and text-to-video retrieval tasks on the MSR-VTT dataset [2]. It comprises 10,000 YouTube video clips and 200,000 captions, each with 20 captions. We adhere to the 1K subset setting during testing for fair comparisons. Performance is evaluated using the $Recall@K$ metric, where $K \in 1, 5, 10$, where K denotes the best-performing K candidates. The retrieval is claimed to be positive if the top K predictions include the paired ground truth.

B. Retrieval Results

As shown in Table I, the proposed VTA method offers competitive performance against benchmarking methods. We only include methods without LLMs in performance benchmarking since we focus on aligning visual and textual data. We use the number of trainable parameters as the metric to show the efficiency claim. Our VTA method only requires 3% trainable parameters, compared to the baseline model, CLIP [1]. Furthermore, VTA requires only six frames in inference. Other methods, which rely on a dense attention mechanism, have a computational cost proportional to the video length.

Regarding interpretability, our VTA method provides intermediate results from genre classification. Genre labels, POS tags, and detected objects can be formulated as conditional probabilities as given in Eq. (2). The retrieval process demands the following tasks: 1) keyword decision in captions and object detection in video, and 2) genre ranking in the subgroups. The VTA alignment is achieved by linking keywords and objects.

C. Ablation Studies

We perform ablation studies to examine the contributions of each module. We divide the whole pipeline into several parts. The vanilla setting uses only the pooled CLIP [1] features as shown in the first row of Table II. Five post-processing steps are considered. They are individually added to the vanilla setting, as reported in columns 2-6 in Table II. The keyframe selection step adopts the six frames from the shot changes. The noun elimination step adopts the co-occurrence matrix between detected objects and noun keywords provided by the POS tags and the object detection results. The verb elimination step exploits the co-occurrence of verbs and detected objects. The verb alignment step uses the verb embedding in the ranking module. After genre detection, the genre clustering step is applied in the subgroup ranking module. Finally, the five steps are added jointly, and the associated three performance metrics are reported in the last column of the table.

V. Conclusion and Future Work

This work proposed a lightweight and transparent video-text alignment (VTA) method. Its state-of-the-art performance in video-to-text and text-to-video retrieval tasks was experimentally demonstrated. Its number of trainable parameters is significantly lower, its inference cost was alleviated via keyframe selection, and its processing time remains constant regardless of the video length. The VTA method is an interpretable stage-wise decision process. The nouns and verbs in the captions and the detected objects in the keyframes conceptualize the visual and textual spaces. Genre classification facilitates ranking for the final decision in a subgroup.

Furthermore, we conducted an ablation study on each element of the VTA method. All elements contribute to the retrieval performance in the MSR-VTT dataset. We plan to develop a purely green learning solution for visual content understanding without leveraging the pre-trained object detector in the future. We will have a more transparent and efficient multimodal learning paradigm if feasible.

Acknowledgment

This work was supported by the DEVCOM Army Research Laboratory (ARL) under agreement W911NF2020157. Computation in the work was supported by the University of Southern California's Center for Advanced Research Computing (carc.usc.edu).

References

[1] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

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

[3] C.-C. J. Kuo and A. M. Madni, "Green learning: Introduction, examples and outlook," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103 685, 2023.

[4] A. Torabi, N. Tandon, and L. Sigal, "Learning language-visual embedding for movie understanding with natural-language," *arXiv preprint arXiv:1609.08124*, 2016.

TABLE I
PERFORMANCE BENCHMARKING OF SEVERAL TEXT-TO-VIDEO AND VIDEO-TO-TEXT RETRIEVAL METHODS FOR THE MSR-VTT DATASET, WHERE THE BEST PERFORMANCE IS LABELED IN **BOLD**, AND THE SECOND BEST IS LABELED IN *italics*.

Method	text-to-video			video-to-text		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
UniVL [17]	21.2	49.6	63.1	-	-	-
HIT-pretrained [18]	30.7	60.9	73.2	32.1	62.7	74.1
CLIP [7]	31.2	53.7	64.2	27.2	51.7	62.6
FROZEN [19]	31.0	59.5	70.5	-	-	-
CLIP4clip [20]	43.1	70.4	80.8	43.1	70.5	81.2
Clip2Video [21]	**45.6**	**72.6**	*81.7*	*43.5*	**72.3**	**82.1**
Ours	*44.5*	*71.4*	**81.7**	**44.0**	*71.2*	*81.6*

TABLE II
ABLATION STUDIES ON THE VTA METHOD.

Method	MSR-VTT		
	Recall@1	Recall@5	Recall@10
Vanilla	24.8	47.2	56.6
+ Keyframe Selection	29.3	54.5	65.0
+ Noun Elimination	32.1	53.6	65.2
+ Verb Elimination	33.3	57.9	69.0
+ Verb Alignment	41.2	68.1	79.2
+ Genre Clustering	42.5	66.9	77.4
Final	44.5	71.4	81.7

[5] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 471–487.

[6] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[7] J. A. Portillo-Quintero, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "A straightforward framework for video retrieval using clip," in *Mexican Conference on Pattern Recognition*, Springer, 2021, pp. 3–12.

[8] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 638–647.

[9] J. Wang, D. Chen, Z. Wu, *et al.*, "Omnivl: One foundation model for image-language and video-language tasks," *Advances in neural information processing systems*, vol. 35, pp. 5696–5710, 2022.

[10] C.-C. J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen, "Interpretable convolutional neural networks via feedforward design," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 346–359, 2019.

[11] Y. Yang, W. Wang, H. Fu, C.-C. J. Kuo, *et al.*, "On supervised feature selection from high dimensional feature spaces," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[12] T.-S. Yang, Y.-C. Wang, C. Wei, S. You, and C.-C. J. Kuo, "Efficient human-object-interaction (ehoi) detection via interaction label coding and conditional decision," *Computer Vision and Image Understanding*, p. 104 390, 2025.

[13] T.-S. Yang, Y.-C. Wang, C. Wei, S. You, and C.-C. J. Kuo, "Gma: Green multi-modal alignment for image-text retrieval," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2024, pp. 1–6.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, Springer, 2020, pp. 213–229.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

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

[17] H. Luo, L. Ji, B. Shi, *et al.*, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

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

[19] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.

[20] H. Luo, L. Ji, M. Zhong, *et al.*, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[21] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2video: Mastering video-text retrieval via image clip," *arXiv preprint arXiv:2106.11097*, 2021.