Date of publication MAR-31, 2024, date of current version FEB-02, 2024.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.1.3440

Attr4Vis: Revisiting Importance of Attribute Classification in Vision-Language Models for Video Recognition

ALEXANDER ZARICHKOVYI[1], INNA V. STETSENKO [1]

[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Prospect Beresteiskyi, Kyiv, 03056, Ukraine

Corresponding author: Alexander Zarichkovyi (e-mail: alexander.zarichkovyi@gmail.com).

ABSTRACT Vision-language models (VLMs), pretrained on expansive datasets containing image-text pairs, have exhibited remarkable transferability across a diverse spectrum of visual tasks. The leveraging of knowledge encoded within these potent VLMs holds significant promise for the advancement of effective video recognition models. A fundamental aspect of pretrained VLMs lies in their ability to establish a crucial bridge between the visual and textual domains. In our pioneering work, we introduce the Attr4Vis framework, dedicated to exploring knowledge transfer between Video and Text modalities to bolster video recognition performance. Central to our contributions is the comprehensive revisitation of Text-to-Video classifier initialization, a critical step that refines the initialization process and streamlines the integration of our framework, particularly within existing Vision-Language Models (VLMs). Furthermore, we emphasize the adoption of dense attribute generation techniques, shedding light on their paramount importance in video analysis. By effectively encoding attribute changes over time, these techniques significantly enhance event representation and recognition within videos. In addition, we introduce an innovative Attribute Enrichment Algorithm aimed at enriching set of attributes by large language models (LLMs) like ChatGPT. Through the seamless integration of these components, Attr4Vis attains a state-of-the-art accuracy of 91.5% on the challenging Kinetics-400 dataset using the InternVideo model.

KEYWORDS computer vision; video recognition; cross-model exploration; vision-language models; lexicon enrichment algorithm.

I. INTRODUCTION

IN recent years, the impressive achievements in large-scale pretraining within the field of Natural Language Processing (NLP) have generated considerable interest within the computer vision community. Notable examples include BERT [1], GPT [2], ERNIE [3], and T5 [4]. These advancements have served as a source of inspiration for researchers, prompting them to explore similar techniques in the domain of computer vision.

Vision-language models (VLMs) are a product of this exploration, harnessing extensive datasets of image-text pairs characterized by weak correspondence and substantial noise for the purpose of contrastive learning. Notable examples in this realm encompass CLIP [5], ALIGN [6], CoCa [7], and Florence [8]. These VLMs have demonstrated remarkable versatility, exhibiting the ability to transfer knowledge effectively across a wide spectrum of visual tasks.

Naturally, this success has given rise to the notion of leveraging the knowledge encoded within these potent, pretrained VLMs as a promising avenue for the development of video recognition models. The current landscape of exploration in this domain can be categorized into several distinct research areas. As illustrated in Fig. 1(a), straightforward approach [9,10] adheres to the conventional unimodal video recognition paradigm by initializing the video encoder with the pretrained visual encoder from a VLM. Conversely, the alternative approach [11-14] directly incorporates the entire VLM into a video-text learning framework, leveraging natural language elements such as class names as supervisory signals, as depicted in Fig. 1(b).

The BIKE framework [15] has brought to light the limitations of prior methodologies, which primarily rely on unidirectional Video-to-Text matching. Such constraints have curtailed the exploitation of the full potential of Vision-Language Models (VLMs) in the context of video recognition. In response, BIKE proposed a novel approach that encourages bidirectional knowledge exploration across the visual and textual domains, as depicted in Fig. 1(c).

The innovative framework introduced in this study pioneers the concepts of Video-to-Text and Text-to-Video knowledge mining. This entails the generation of textual information from

the input video, a crucial process known as attribute generation. Furthermore, BIKE [15] harnesses category descriptions to establish temporal saliency, thereby facilitating the extraction of valuable video-related signals. Through these mechanisms, BIKE [15] lays the foundation for comprehensive knowledge exploration across visual and textual modalities, thereby enhancing video understanding and recognition capabilities across various domains and tasks.

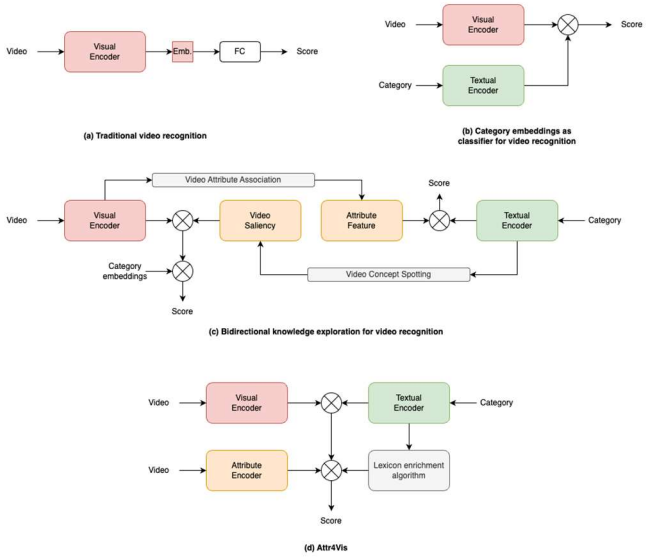

Figure 1. Illustration of the difference between our Attr4Vis paradigm (d) with existing unimodal paradigm (a), cross-modal paradigm (b) and bi-directional paradigm (c).

In our research, we underscore the critical importance of unsupervised attribute predictions in enhancing video recognition. In alignment with this objective, our aim is to streamline the BIKE framework by eliminating the reliance on temporal saliency within the Text-to-Video branch. Instead, we advocate for the integration of a single classifier with Text-to-Video pretraining transfer, following the principles of Text4Vis [14]. Moreover, we enhance the process of Video-to-Text transfer by conducting attribute classification on densely overlapping segments of videos. This innovative approach allows us to encode information regarding attribute changes over time, thereby enhancing the representation of events within video data through the utilization of Large Language Models (LLMs). Furthermore, we underscore the significance of a predefined lexicon for attribute detection by VLMs. We demonstrate that this lexicon can be enriched through conversational LLMs, such as ChatGPT, which enables an expanded range of attribute recognition possibilities within video data.

The Attr4Vis framework pioneers the exploration of knowledge transfer between Video and Text modalities to enhance video recognition performance by revisiting Text-to-Video classifier initialization, emphasizing dense attribute generation techniques, and introducing an Attribute Enrichment Algorithm. By seamlessly integrating these components, Attr4Vis advances the state of the art in video recognition, offering promising avenues for knowledge transfer and improved attribute recognition capabilities.

Our contributions in this work encompass several key aspects:

• **Revisiting Text-to-Video Classifier Initialization**: We revisit the crucial aspect of initializing the Text-to-Video classifier, a step that allows us to remove Video Concept Spotting block which plays a pivotal role in streamlining the BIKE framework [15]. This initiative facilitates a more seamless integration of our proposed methodology into existing VLMs.

• **Emphasizing the Significance of Dense Attribute Generation**: We underscore the paramount importance of dense attribute generation techniques in the context of video analysis. This approach allows for the effective encoding of attribute changes over time, contributing to enhanced event representation and recognition.

• **Introducing Attribute Enrichment Algorithm**: We introduce novel algorithm aimed at enriching the set of attributes associated with video data. This augmentation of attributes offers an expanded spectrum of possibilities for attribute recognition within video datasets, thereby broadening the applicability of our framework.

Our framework, Attr4Vis, synthesizes these contributions to advance the state of the art in video recognition, offering promising avenues for the exploration of knowledge transfer between visual and textual domains, as well as improved attribute recognition capabilities.

II. RELATED WORK

A. IMAGE-LANGUAGE MODELS

Visual recognition has conventionally relied on convolutional neural networks (CNNs) as the primary backbone architecture for image and video recognition tasks [16, 17]. However, the success of the Transformer architecture in Natural Language Processing, notably exemplified by [18], inspired the development of the Vision Transformer (ViT) [19], which directly applies the Transformer to images, yielding remarkable performance gains in image recognition. Consequently, ViT [19] has initiated a paradigm shift in image recognition backbones, transitioning from CNNs to Transformers. Subsequent studies, such as DeiT [20] and Swin [21], have emerged to further enhance performance. Moreover, the application of Transformers to video recognition has gained momentum with the introduction of models like TimeSFormer [22], ViViT [23], VideoSwin [24], and MViT [25].

In the domain of image-language pretraining, the CLIP model [5] has emerged as a benchmark for coordinated vision-language pretraining. CLIP utilizes the image-text InfoNCE contrastive loss [26] and has spawned several variants [7] that combine various learning tasks, including image-text matching and masked image/language modeling. These contrastively learned models exhibit two noteworthy characteristics: rich visual feature representations and aligned textual feature representations. Another study [27] incorporated the downstream classification task into the pretraining phase, resulting in a significant improvement in accuracy compared to standard cross-entropy loss.

In the context of video-text learning, several methods [28, 29] have harnessed vision-language pretraining for video-text retrieval. Furthermore, recent approaches [11] extend the CLIP model [5] to train downstream video-text matching models using contrastive loss and subsequently employ the similarity between learned video and text embeddings for video recognition during inference.

In contrast to the contrastive-based methods mentioned above, our approach investigates the efficient feature transfer

between visual and textual domain within the standard visual recognition paradigm. We directly extract both visual and textual knowledge from the video to enable better utilization of foundational models.

B. CROSS-MODAL FOUNDATION MODELS

Foundation models in the realm of computer vision have typically been tailored to specific tasks and domains, often requiring manual annotation of datasets for their training. However, recent research has introduced the concept of vision foundation models aimed at alleviating these constraints. Notably, CLIP [5] and ALIGN [6] have harnessed vast web-scale collections of noisy image-text pairs to train dual-encoder models using contrastive learning. This approach yields robust image-text representations, enabling powerful zero-shot transfer capabilities. INTERN [30] extends this paradigm by incorporating multiple stages of self-supervised pretraining, leveraging a substantial quantity of image-text pairs alongside manually annotated images. INTERN [30] exhibits superior linear probe performance compared to CLIP [5] and enhances data efficiency in downstream image tasks.

Florence [8] further advanced this line of research by integrating unified contrastive learning [31] and sophisticated adaptation models, thereby facilitating a wide array of vision tasks across diverse transfer settings. SimVLM [32] and OFA [33], on the other hand, pursue encoder-decoder model training with generative targets, delivering competitive performance across various multimodal tasks. Additionally, CoCa [7] unifies contrastive learning akin to CLIP [5] with generative learning akin to SimVLM [32]. Notably, Beit-3 [34] introduces Multiway Transformers within the framework of unified BeiT [35] pretraining, achieving state-of-the-art transfer results across multiple vision and image-language tasks.

In the domain of video foundation models, prior models like those exemplified by CoCa [7] and Florence [8] have primarily excelled in video recognition, particularly in datasets such as Kinetics. However, when it comes to multimodal tasks involving video, models such as VIOLET [36] leverage masked language and masked video modeling, All-in-one [37] proposed unified video-language pretraining with a shared backbone, and LAVENDER [38] unified tasks through masked language modeling. While these models perform admirably in multimodal benchmarks, their training data often remains limited with regard to video-text pairs, thereby struggling when applied to video-only tasks, such as action recognition.

In contrast, MERLOT Reserve [39] breaks new ground by amassing a vast collection of 20 million video-text-audio pairs for joint video representation training, employing contrastive span matching. Consequently, it achieves a state-of-the-art performance not only in video recognition but also in visual commonsense reasoning. It is worth noting that current video foundation models, when compared to their image foundation counterparts, exhibit limitations in their support for video and video-language tasks, especially in the context of fine-grained temporal discrimination tasks such as temporal localization.

The paradigm of multimodal pretraining has become a cornerstone in the video-language domain, beginning with the development of image-text pretraining and evolving into large-scale video-text pretraining with subsequent fine-tuning for specific downstream tasks [36, 40, 41]. Seminal methods [42, 43] have traditionally leveraged pretrained visual and language encoders to extract offline video and text features. More recent approaches [37, 39], however, have demonstrated the feasibility of end-to-end training, streamlining the process. These methods often encompass two or three pretraining tasks, including masked language modeling [38], video-text matching [37], video-text contrastive learning [41], and video-text masked modeling [36], among others.

In contrast to the previous work, InternVideo [30] stands as a versatile video foundation model, accompanied by its training regimen and intrinsic collaborative mechanisms. In terms of architectural design, InternVideo [30] adopts the Vision Transformer (ViT) [19] as its foundational structure, augmented by the inclusion of UniformerV2 [30] and additional localized spatiotemporal modeling modules. This amalgamation facilitates the creation of multi-level representations with robust interaction capabilities.

In the realm of learning, InternVideo [30] progressively enhances its representation by seamlessly integrating both self-supervised techniques, encompassing masked modeling and multimodal learning, alongside supervised training. Moreover, InternVideo [30] dynamically derives novel features from these two transformers through learnable interactions, thereby harnessing the strengths of generative and contrastive learning paradigms. The culmination of these efforts yields a remarkable outcome, as InternVideo [30] established new performance benchmarks across 34 datasets spanning 10 prominent video-related tasks.

In contrast to the cross-modal foundation models training, our approach focuses on utilization of these foundation models in a way that will yield the best performance.

III. METHODOLOGY

An overview of our proposed Attr4Vis is shown in Fig. 2. Next, we will elaborate on each component in more detail.

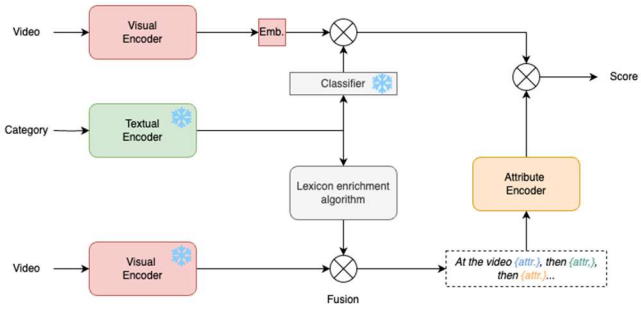

Figure 2. An overview of our Attr4Vis for video recognition

A. TEXT-TO-VIDEO BRANCH: TEXTUAL EMBEDDINGS VECTORS

For Text-to-Video knowledge exploration we follow textual embedding encoding from [30] to transfer semantic knowledge from text to visual model. It builds projection weights W composed of the embedded textual feature vectors of dataset labels L. Given a set of tokenized class labels $L = \{l_1, l_2, l_3, \ldots, l_c\}$, we have:

$$W_i \sim TextEcoder(l_i), i = 1,2,3,\ldots,c, \quad (1)$$

where W_i the i-th row vector in matrix W. Hence W_i is initialized using LLM output of textual label of the i-th class. As TextEncoder in our experiments we used Multilingual-E5-large [31].

B. Text-to-Video branch: Textual Embeddings Vectors

As illustrated in Fig. 3, we leverage the zero-shot capabilities inherent in Vision-Language Models (VLMs), exemplified by InternVideo [30], to discern the most pertinent phases within a predefined lexicon and designate them as potential "Attributes" for the given video content. This process unfolds as follows: we commence by applying image encoder to the small chunks of the input video, thereby extracting frame-level features for each chunk, and subsequently aggregating using average pooling that yields a comprehensive chunk embedding.

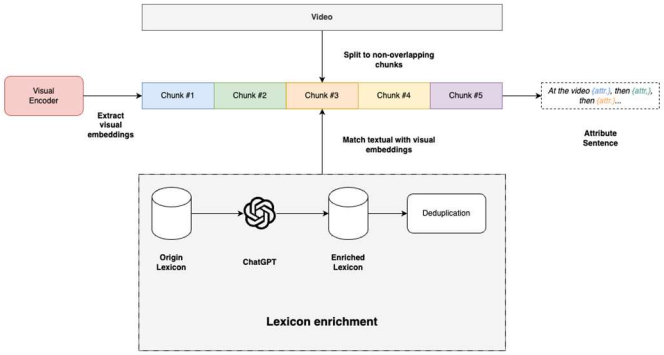

Figure 3. An overview of Video-to-Text branch. It generates Attribute sentence for given input video and predefined lexicon. Attribute generation happens for non-overlapping chunks by matching visual embeddings with textual.

Simultaneously, we feed predefined lexicon into text encoder, resulting in the generation of a set of text embeddings. We proceed to compute the similarity between each chunk embedding and text embeddings, meticulously sorting the outcomes. Subsequently, we select the top few phrases that exhibit the highest degree of alignment as the designated "Attributes" for the chunk.

Having successfully identified these attributes, we employ a straightforward fusion technique that concatenates them into a unified attributes sentence. To enhance interpretability and context, we append a manually designed prompt as a prefix to this sentence, typically in the form of "At the video {}, then {}, then {}..." This comprehensive approach ensures the creation of meaningful and contextually relevant auxiliary attributes for our video content analysis.

C. TEXT-TO-VIDEO BRANCH: TEXTUAL EMBEDDINGS VECTORS

Traditionally, video recognition datasets have been constrained by a limited number of categories defined by the dataset labels. In the context of the BIKE framework [15], the original dataset labels were exclusively employed for the creation of a lexicon. However, it is evident, as demonstrated in our ablation study (refer to Section 5.2), that relying solely on this original set of labels imposes restrictions on the capacity of Vision-Language Models (VLMs), exemplified by InternVideo [30], to generate comprehensive prompts for the Video-to-Text branch.

To address this limitation and broaden the variability of available labels, we propose an innovative approach. Specifically, we advocate for the augmentation of the lexicon by harnessing conversational LLMs, with a specific focus on ChatGPT. Through this method, we significantly enrich the lexicon at the disposal of VLMs. The detailed formalization of our lexicon enrichment procedure is provided in Algorithm 1. This augmentation strategy not only enhances the descriptive capabilities of VLMs but also facilitates more nuanced and contextually relevant prompts, thus contributing to improved video understanding and recognition performance.

Algorithm 1. Proposed algorithm for lexicon enrichment

Inputs: Dataset D, set dataset labels $L = \{L_1, L_2, \dots, L_n\}$, Textual Embedding model M, deduplication threshold τ.

// Generate enriched labels

1. $L_e = \{\}$ *// Initialize set of enriched labels*
2. For each L_i in L:
   // Preparing textual prompt for each label
   2.1. prompt = "You are a domain expert in the $\{D\}$ dataset, helping develop a labeling system. Generate additional labels that will enrich $\{L_i\}$ label of $\{D\}$ dataset. Format output as comma separated labels."
   2.2. result = ChatGPT(prompt) *// Generate additional labels*
   2.3. $L_e.add(result)$

// Deduplicate labels by semantic meaning

3. $E_e = \{M(L_i)\ for\ L_i\ in\ L_e\}$ // Generate embeddings for each label
4. $L_o = \{\}$ *// Initialize output labels*
5. For e in E_e:
   5.1. distance = cosine(e, L_o) // Calculate cosine distance for each pair
   5.2. if all(distance) < τ:
      5.2.1. $L_o.add(e)$

IV. EXPERIMENTS AND RESULTS

A. SETUP

Our experimental investigations are conducted on Kinetics-400 [46] dataset.

We adopt the visual encoder component of InternVideo [30] as the foundation for our video encoder. Similarly, the textual encoder of InternVideo [30] is utilized for attributes encoder. To mitigate potential conflicts between these two branches, we employ a sequential training approach, first training the video encoder and subsequently addressing the attributes encoder. In the preparation of video inputs, we judiciously employ sparse sampling techniques involving a variable number of frames denoted as T (e.g., 8, 16, 32). In all our experiments we used $\tau = 0.85$ for deduplicating enriched attribute lexicon. For attribute prediction branch we sampled each video into 8 non-overlapping temporal chucks with 16 frames in each (section 5.1) and used prediction with the biggest score to represent this chunk.

Our evaluation protocol embrace the "Multiple Views" strategy, a common practice in the field [6, 12, 40], which entails the sampling of multiple clips per video along with several spatial crops to achieve higher accuracy. For the purpose of benchmarking and comparisons with state-of-the-art approaches, we specifically employ a configuration involving four clips with three crops, denoted as "4×3 Views," as detailed in Table 1.

B. RESULTS

We present our results on Kinetics-400 in Table 1 and compare our approach with SOTAs trained under various pretraining settings. Our approach outperforms regular video recognition methods while requiring significantly less computation.

Table 1. Comparisons with state-of-the-art methods on Kinetics-400. We report the FLOPs in inference phase. "Views" indicates # temporal clip × # spatial crop. The magnitudes are Giga (10^9) and Mega (10^6) for FLOPs and Params.

Method	Input	Top-1 (%)	Top-5 (%)	Views	FLOPs	Params
NL I3D-101 [47]	128 × 224²	77.7	93.3	10×3	359 × 30	61.8
MVFNet [48]	24 × 224²	79.1	93.8	10×3	188 × 30	-
TimeSformer-L [49]	96 × 224²	80.7	94.7	1×3	2380 × 3	121
ViViT-L/16×2 [23]	32 × 320²	81.3	94.7	4×3	3992 × 12	311
VideoSwin-L [24]	32 × 384²	84.9	96.7	10×5	2107 × 50	200
Methods with large-scale image pre-training						
ViViT-L/16×2 [23]	32 × 320²	83.5	95.5	4×3	3992 × 12	311
ViViT-H/16×2 [23]	32 × 224²	84.8	95.8	4×3	8316 × 12	648
TokenLearner-L/10 [50]	32 × 224²	85.4	96.3	4×3	4076 × 12	450
MTV-H [51]	32 × 224²	85.8	96.6	4×3	3706×12	-
CoVeR [52]	16 × 448²	87.2	-	1×3	-	-
Methods with large-scale image-language pre-training						
CoCa ViT-giant [7]	6 × 288²	88.9	-	-	-	2100
VideoPrompt ViT-B/16 [11]	16 × 224²	76.9	93.5	-	-	-
ActionCLIP ViT-B/16 [13]	32 × 224²	83.8	96.2	10×3	563 × 30	142
Florence [8]	32 × 384²	86.5	97.3	4×3	-	647
ST-Adapter ViT-L/14 [10]	32 × 224²	87.2	97.6	3×1	8248	-
EVL ViT-L/14 [9]	32 × 224²	87.3	-	3×1	8088	-
X-CLIP ViT-L/14 [12]	16 × 336²	87.7	97.4	4×3	3086 × 12	-
Text4Vis ViT-L/14 [14]	32 × 336²	87.8	97.6	1×3	3829 × 3	230
BIKE ViT-L/14 [15]	32 × 336²	88.6	98.3	4×3	3728 × 12	230
InternVideo-T [30]	32 × 336²	91.1	99.0	4×3	1434 × 12	1300
Attr4Vis (ours)	**32 × 336²**	**91.5**	**99.2**	**4×3**	**1434 × 12**	**1300**

V. ABLATION STUDY

In the ablation study section, our objective is threefold. Firstly, we aim to evaluate the impact of dense attribute prediction on model performance. Secondly, we aim study importance of lexicon enrichment algorithm. Lastly, we studied aspects of initializing the Text-to-Video classifier and removing Video Concept Spotting branch of BIKE [15]. This comparative analysis will provide insights into the efficacy of our approach and its potential to outperform the state-of-the-art technique. Additionally, by isolating the contribution of each change, we can better understand the relative importance of each modification and its impact on the overall performance of the solution.

A. STUDY THE SIGNIFICANCE OF DENSE ATTRIBUTE DETECTION

Our experiential study (Table 2) underscores the paramount importance of dense attribute generation techniques in the context of video analysis. Reasoning for such improvements is that approach allows for the effective encoding of attribute changes over time with high granularity, contributing to the enhanced event representation and recognition.

Table 2. Impact of dense attribute prediction on the performance. "Views" indicates #non-overlapping temporal chunks × # frames.

Method	Views	Top-1 (%)
Without attribute prediction	-	91.1%
Sparse attribute prediction	1x16	91.3% (+0.2%)
Dense attribute prediction	4x16	91.4% (+0.3%)
Dense attribute prediction	8x16	91.5% (+0.4%)
Dense attribute prediction	16x16	91.5% (+0.4%)

B. STUDY ON IMPORTANCE OF LEXICON ENRICHMENT

We compared our proposed algorithm for lexicon enrichment (section 3.3) with ImageNet and Kinetics-400 lexicons in Table 3.

Table 3. Impact of different lexicons on performance

Lexicon	Top-1 (%)
Without attribute prediction	91.1%
ImageNet-1K	90.3% (-0.8%)
Kinetics-400	91.4% (+1.1%)
Lexicon enrichment algorithm	91.5% (+0.1%)

C. STUDY ON IMPORTANCE OF LEXICON ENRICHMENT

We revisit the crucial aspect of initializing the Text-to-Video classifier, a step that plays a pivotal role in streamlining the BIKE framework [15]. This makes it possible to drop Video Concept Spotting branch hence streamlining integration of our proposed methodology into existing VLMs. The results of the study are available in Table 4.

Table 4. Impact of initialization and Video Concept Spotting branch on the performance.

Change	Top-1 (%)
BIKE [15]	88.6%
+ InternVideo-T [29]	90.6% (+2.0%)
+ Text4Vis initialization [14]	90.8% (+0.2%)
- Video Concept Spotting branch	91.2% (+0.4%)
+ Dense attribute prediction	91.4% (+0.2%)
+ Lexicon enrichment	91.5% (+0.1%)

VI. CONCLUSION

In this study, we present an innovative framework known as Attr4Vis, aimed at exploring the transfer of knowledge between video and text modalities to enhance video recognition. Our contributions in this work encompass several significant components. Firstly, we underscore the paramount importance of employing dense attribute generation techniques within the realm of video analysis. This approach allows for the effective encoding of attribute changes over time, thereby contributing to improved event representation and recognition. Secondly, we introduce a novel algorithm designed to enrich the array of attributes associated with video data. This augmentation of attributes broadens the spectrum of possibilities for attribute recognition within video datasets, thereby enhancing the applicability of our framework. Furthermore, we revisit the critical aspect of initializing the Text-to-Video classifier and removing Video Concept Spotting branch, a pivotal step that streamlines the BIKE framework [15].

Our framework, Attr4Vis, amalgamates these contributions, advancing the forefront of video recognition achieving state-of-the-art results of 91.5% Top-1 accuracy on Kinetics-400 dataset. It offers promising opportunities for exploring knowledge transfer between the visual and textual domains, and concurrently enhances attribute recognition capabilities.

References

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, et al., "Language models are few-shot learners," in: H. Larochelle and M. Ranzato and R. Hadsell and M.F. Balcan and H. Lin (Eds.), *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, 2020, pp. 1877–1901.

[3] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "Ernie: Enhanced

language representation with informative entities," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 28 July – 2 August 2019, pp. 1441–1451. https://doi.org/10.18653/v1/P19-1139.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, et al., "Learning transferable visual models from natural language supervision," *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, Virtual, 18-24 July 2021, pp. 8748–8763.

[6] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, et al., "Scaling up visual and vision-language representation learning with noisy text supervision," *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, Virtual, 18-24 July 2021, pp. 4904–4916.

[7] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini and Y. Wu. *Coca: Contrastive captioners are image-text foundation models*, 2022, [Online]. Available at: https://arxiv.org/abs/2205.01917.

[8] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, et al., *Florence: A new foundation model for computer vision*, 2021, [Online]. Available at: https://arxiv.org/abs/2111.11432.

[9] Z. Lin, S. Geng, R. Zhang, P. Gao, et al., "Frozen CLIP models are efficient video learners," *Lecture Notes in Computer Science*, vol. 13695, pp. 388–404, 2022. https://doi.org/10.1007/978-3-031-19833-5_23.

[10] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "St-adapter: Parameter-efficient image-to-video transfer learning," S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh Advances (Eds.), *Advances in Neural Information Processing Systems (NeurIPS'2022)*, vol. 35, 2022, pp. 26462-26477.

[11] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," *Lecture Notes in Computer Science*, vol. 39, 2022, pp. 105–124. https://doi.org/10.1007/978-3-031-19833-5_7.

[12] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," *Proceedings of the Computer Vision–ECCV 2022: 17th European Conference*, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV. Springer, 2022, pp. 1–18. https://doi.org/10.1007/978-3-031-19772-7_1.

[13] M. Wang, J. Xing, and Y. Liu. *Actionclip: A new paradigm for video action recognition*, 2021, [Online] Available at: https://arxiv.org/abs/2109.08472.

[14] W. Wu, Z. Sun, and W. Ouyang, "Revisiting classifier: Transferring vision-language models for video recognition," *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, Washington, D.C., USA, February 7–14, 2023, pp. 2847-2855. https://doi.org/10.1609/aaai.v37i3.25386.

[15] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang, "Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023,* Vancouver, BC, Canada, June 17-24, 2023, pp. 6620-6630. https://doi.org/10.1109/CVPR52729.2023.00640.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proceedings of the International Conference NeurIPS*, 2012, pp. 1-9.

[17] I. Paliy, A. Sachenko, V. Koval and Y. Kurylyak, "Approach to face recognition using neural networks," Proceedings of the *2005 IEEE Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, IDAACS'2005, Sofia, Bulgaria, 2005, pp. 112-115, https://doi.org/10.1109/IDAACS.2005.282951.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., *An Image is Worth 16x16 words: Transformers for Image Recognition at Scale*, 2020, [Online]. Available at: https://arxiv.org/abs/2010.11929.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, *Training Data-efficient Image Transformers & Distillation Through Attention*, 2020, [Online]. Available at: https://arxiv.org/abs/2012.12877.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, et el., "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, Montreal, QC, Canada, October 10-17, 2021, pp. 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986.

[22] G. Bertasius, H. Wang, and L. Torresani, "Is SpaceTime Attention All You Need for Video Understanding?" *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, Virtual, 18-24 July, 2021, pp. 813–824.

[23] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C.V. Schmid, "ViViT: A video vision transformer," *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, Montreal, QC, Canada, October 10-17, 2021, pp. 6816-6826. https://doi.org/10.1109/ICCV48922.2021.00676.

[24] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin and H. Hu, "Video swin transformer," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, New Orleans, LA, USA, June 18-24, 2022, pp. 3202–3211. https://doi.org/10.1109/CVPR52688.2022.00320.

[25] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, et el., "Multiscale vision transformers," *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, Montreal, QC, Canada, October 10-17, 2021, 6804-6815. https://doi.org/10.1109/ICCV48922.2021.00675.

[26] A. Van den Oord, Y. Li, and O. Vinyals, *Representation Learning with Contrastive Predictive Coding*, 2018, [Online]. Available at: https://arxiv.org/abs/1807.

[27] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, New Orleans, LA, USA, June 18-24, 2022, pp. 19141-19151. https://doi.org/10.1109/CVPR52688.2022.01857.

[28] L. Wang, Z. Tong, B. Ji, and G.Wu, "TDN: Temporal difference networks for efficient action recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, Virtual, June 19-25, 2021, pp. 1895–1904. https://doi.org/10.1109/CVPR46437.2021.00193.

[29] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022. https://doi.org/10.1016/j.neucom.2022.07.028.

[30] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, et el., *InternVideo: General Video Foundation Models via Generative and Discriminative Learning*, 2022, [Online]. Available at: https://arxiv.org/abs/2212.03191.

[31] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, et el., *Text Embeddings by Weakly-Supervised Contrastive Pre-training*, 2022, [Online]. Available at: https://arxiv.org/abs/2212.03533.

[32] Z. Wang, J. Yu, A.W. Yu, Z. Dai, Y. Tsvetkov and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," *Proceedings of the Tenth International Conference on Learning Representations, ICLR 2022*, Virtual Event, April 25-29, 2022.

[33] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, et el., "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," *Proceedings of the 39th International Conference on Machine Learning*, *ICML 2022*, Baltimore, Maryland, USA, July 17-23, 2022, pp. 23318-23340.

[34] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, et al., *Image as a Foreign Language: Beit Pretraining for all Vision and Vision-language Tasks*, 2022. https://doi.org/10.1109/CVPR52729.2023.01838.

[35] H. Bao, L. Dong, S. Piao, F. Wei, "BEiT: BERT pre-training of image transformers," *Proceedings of the Tenth International Conference on Learning Representations, ICLR 2022*, Virtual Event, April 25-29, 2022.

[36] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Yang Wang, et el., *Violet: End-to-end video-language transformers with masked visual-token modeling*, 2021, [Online]. Available at: https://arxiv.org/abs/2111.12681.

[37] A. J. Wang, Y. Ge, R. Yan, Y. Ge, X. Lin, et el., *All in one: Exploring unified video-language pre-training*, 2022. https://doi.org/10.1109/CVPR52729.2023.00638.

[38] L. Li, Z. Gan, K. Lin, C.-C. Lin, Z. Liu, C. Liu and L. Wang, *Lavender: Unifying video-language understanding as masked language modeling*, 2022. https://doi.org/10.1109/CVPR52729.2023.02214.

[39] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, et el., "MERLOT RESERVE: Neural script knowledge through vision and language and sound," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, New Orleans, LA, USA, June 18-24, 2022, pp. 16354-16366. https://doi.org/10.1109/CVPR52688.2022.01589.

[40] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, Seattle, WA, USA, June 13-19, 2020, pp. 9876-9886. https://doi.org/10.1109/CVPR42600.2020.00990.

[41] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, *Videoclip: Contrastive Pre-training for Zero-shot Video-text Understanding*, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.544.

[42] C. Sun, A. Myers, C. Vondrick, K. P. Murphy, and C. Schmid. "Videobert:

A joint model for video and language representation learning," *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, Seoul, Korea (South), October 27 - November 2, 2019, pp. 7463-7472. https://doi.org/10.1109/ICCV.2019.00756.

[43] L. Zhu, Y. Yang, "ActBERT: Learning global-local video-text representations," *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, Seattle, WA, USA, June 13-19, 2020, pp. 8743-8752. https://doi.org/10.1109/CVPR42600.2020.00877.

[44] J. Lei, L. Li, L. Zhou, Z. Gan, T.L. Berg, M. Bansal and J. Liu. "Less is more: ClipBERT for video-and-language learning via sparse sampling," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, Virtual, June 19-25, 2021, 2021, pp. 7331-7341. https://doi.org/10.1109/CVPR46437.2021.00725.

[45] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: a joint video and image encoder for end-to-end retrieval," *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, Montreal, QC, Canada, October 10-17, 202, pp. 1708-1718. https://doi.org/10.1109/ICCV48922.2021.00175.

[46] W. Kay, J. Carreira, K. Simonyan, B. Zhang, et al. *The Kinetics Human Action Video Dataset,* 2017, [Online]. Available at: https://arxiv.org/abs/1705.06950.

[47] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, *A Short Note about Kinetics600*, 2018, [Online]. Available at: https://arxiv.org/abs/1808.01340.

[48] F. C. Heilbron, V. Escorcia, B. Ghanem, J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015,* Boston, MA, USA, June 7-12, 2015, pp. 961-970. https://doi.org/10.1109/CVPR.2015.7298698.

[49] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, Virtual, 18-24 July, 2021, pp. 813–824.

[50] M. Ryoo, A.J. Piergiovanni, A. Arnab, M. Dehghani, A. Angelova. "TokenLearner: adaptive space-time tokenization for videos," *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, December 6-14, 2021, virtual, pp. 12786–12797.

[51] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun and C. Schmid, "Multiview transformers for video recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, New Orleans, LA, USA, June 18-24, 2022, pp. 3333–3343, 2022. https://doi.org/10.1109/CVPR52688.2022.00333.

[52] B. Zhang, J. Yu, C. Fifty, W. Han, A.M. Dai, R. Pang, and F. Sha, *Co-training Transformer with Videos and Images Improves Action Recognition*, 2021, [Online] Available at: https://arxiv.org/abs/2112.07175.

ALEXANDER ZARICHKOVYI. ***Kaggle competition master, PhD. Student at Igor Sikorsky Kyiv Polytechnic Institute. Specializing on research in computer vision field and video recognition.***

INNA V. STETSENKO ***Doctor of Science, Professor of the Department of Computer Science and Software Engineering, Igor Sikorsky Kyiv Polytechnic Institute. Research interests include parallel computing, artificial intelligence, simulation, and Petri nets.***