Dissecting Video-Language Alignment at Event Level

Anonymous ACL submission

Abstract

Alignment between video and text data is crucial in video-language representation learning and serves as an effective pre-training objective for downstream tasks. Recent video-language models have demonstrated proficiency in learning expressive representations from a large corpus of web-crawled video-text pairs in a contrastive manner by enforcing alignment between videos and their paired caption. However, videos often encapsulate multiple events with a dynamic nature, whereas paired textual data usually contains only broad-level and partial information. This leads to semantic discrepancies in video and text data in such datasets. We question whether video and language can still be well-aligned when trained with such video-text paired data at a finer granularity, as event level. In this work, we propose a novel systematic framework to dissect video-language alignments at event level. Considering videos' dynamic and complex nature, we investigate three types of alignments: temporal, neighborhood, and factual. Our framework proposes extracting multi-aspect events in videos from video scene graph annotations and introduces an alignment score to measure alignment between a video and events. We thoroughly evaluate six existing video-language models on two datasets with eight fine-grained alignment categories and reveal that video-language models demonstrate poor event-level alignment, especially when aligning videos with temporal ordering or attribute binding of a single event, irrespective of model architecture or data scale. In response, we propose incorporating misaligned events as hard negative samples in training video-language models, yielding better alignment in multiple aspects and improved performance in downstream tasks.

1 Introduction

With an influx of multi-media data uploaded on the Internet, videos and texts have become pivotal data modalities in multi-modal learning. This data encapsulates a huge general human knowledge base and constitutes a perfect multi-modal corpus for achieving artificial general intelligence.

Video-language models are data-driven methods that attempt to leverage large-scale video-text data to learn a joint video-language representation(Wang et al., 2022; Miech et al., 2020; Bain et al., 2021a; Li et al., 2022a; Luo et al., 2022; Lei et al., 2022; Zeng et al., 2023; Bansal et al., 2023a) by enforcing higher semantic similarity between video-text pairs and lower similarity for unrelated videos and texts. Previous research(Wang et al., 2022; Jia et al., 2021; Radford et al., 2021) has shown that leveraging web-crawled video-text pairs within such a contrastive learning framework enables exceptional capacities for learning good semantic video representations, bypassing extensive human annotations and curation and establishing a prevalent paradigm for multi-modal representation learning. By aligning videos and texts in a unified representation space, these models have shown further significant advancement in downstream video understanding tasks like Video Question Answering(Li et al., 2023a; Wang et al., 2023c) and Video Retrieval(Luo et al., 2022; Lei et al., 2022; Wang et al., 2023c).

However, *blindly* aligning holistic video and text pairs brings some unsettling concerns: videos and texts are heterogeneous in their structure and semantics: videos usually contain multiple events involving a dynamic structure with time-evolving relations and attributes, whereas texts, particularly crowdsourced captions web-crawled in an uncurated way, usually refer to partial video contents or describe videos at a broad level. This constitutes merely weak supervision.

This raises a question: are video and language well-aligned at a finer granularity after we enforce alignment between holistic videos and texts via training on such datasets? Precisely, are video representations well-aligned with specific and diverse

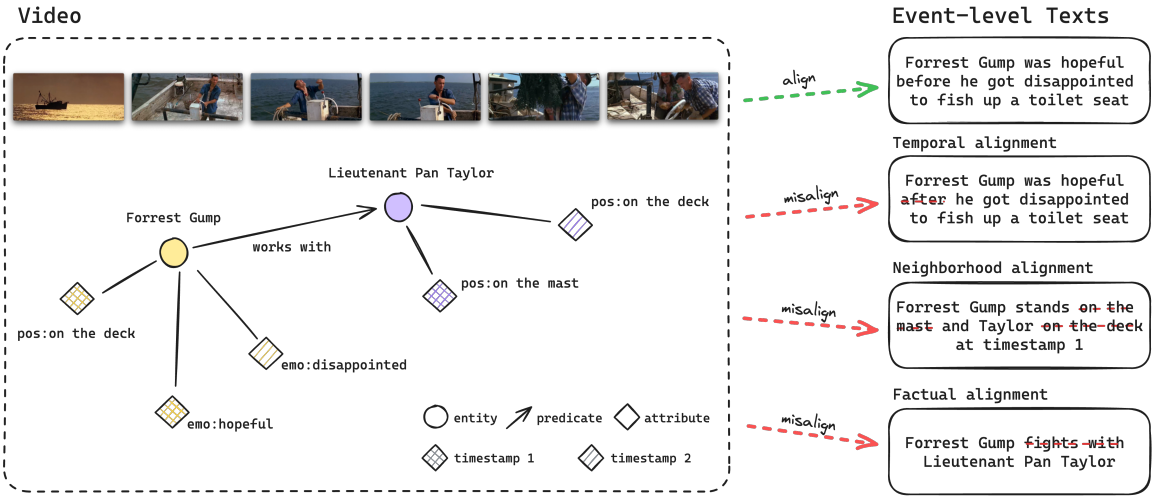

Figure 1: **Video-language alignment at event level.** Web-crawled video-text pairs ignore the semantic and structure discrepancy of cross-modal data and usually lead to weak supervision; video-language models are trained on such datasets to align holistic videos and broad-level textual captions; this makes us wonder *whether video-language are aligned at event level,* with any specific and diverse events?

events as texts, which are fundamental elements of video content that only describe partial content?

To answer this question, we attempt to dissect video-language model alignment *at event level*. As shown in Fig. 1, a holistic video contains multiple events, each of which is spanned by a tuple formulated as "<*S*ubject, *S*ubject *a*ttribute, *P*redicate, *O*bject, *O*bject *a*ttribute, *T*ime*s*tamp>", (s, sa, p, o, oa, ts). The challenge lies in the fact that when pre-trained with video-text pairs, video-language models are expected to align a video with diverse textual captions that describe events in a semantic space. They must also distinguish events inconsistent with video content, even with only minor differences. As illustrated in Fig. 1, such a video should be correctly alignment with a truthful event as "*Forrest Gump was hopeful then disappointed*", and also should distinguish itself from other dubious events, such as "*Forrest Gump was disappointed then hopeful*" with incorrect temporal information, "*Forrest Gump stands on the mast*", with wrong attribute, and "*Forrest Gump fights with Taylor*", with totally unfactual details.

In this work, we propose a novel framework to evaluate existing video-language models on event-level video-language alignment systematically. We define an alignment score that reflects video-language alignment consistency. According to the videos' dynamic and complex nature, we propose three types of event-level alignments: temporal, neighborhood, and factual alignment, corresponding to the example shown in Fig. 1. Temporal alignment focuses on consistent temporal ordering, neighborhood alignment emphasizes distinguishing involved entities in a video, and factual alignment emphasizes understanding word-level knowledge in a video. Correspondingly, we define three manipulation methods according to the alignment types and targetfully generate misaligned events. Our aim is to assess the ability of video-language models to differentiate truthful and manipulated events semantically, thereby pinpointing event-level video-language alignment. Specifically, in order to extract diverse events from a video and manipulate them controllably, we propose extracting events from video scene graph annotations.

We evaluate six video-language models on two datasets and conduct an in-depth analysis. Our results reveal that video-language models demonstrate poor event-level alignment across models and training strategies. Videos are poorly aligned with text when they contain dynamic events or involve multiple confounding entities. For factual alignment, most video-language models show relatively better alignment with word-level knowledge.

To mitigate the issues of poor event-level alignment, we propose to improve it by utilizing misaligned events as hard negative samples for post-training video-language models. Our experimental studies show that augmenting with a mixture of additional hard negative samples can improve the neighborhood and actual alignment at the same time and also benefit downstream tasks like Video Question Answering. Nonetheless, by contrast,

enhancing temporal alignment remains a persistent challenge. This underscores the importance of advanced model architectures for video-language models, especially video encoders with superior temporal modeling, to achieve improved event-level alignment.

To summarize, our contributions include:

- We propose a new framework based on dynamic video scene graph annotation to systematically dissect video-language alignment at event level for three types of alignment: temporal, neighborhood, and factual alignment;
- We benchmark six existing video-language models on two datasets and conduct an in-depth analysis of video-language models' event-level alignment;
- We propose using misaligned events as hard negative samples for post-training to improve multi-aspect video-language alignment.

2 Event-level Video-Language Alignment

In this section, we start by giving a formal definition of events, our metrics to evaluate the event-level alignment of video-language models, and defining three manipulation methods.

2.1 Preliminary

Video Scene Graphs Video scene graphs $\mathcal{G}$ are a concise and holistic semantic representation of video content. Considering the dynamic and complex nature of videos, video scene graphs are temporal and multi-relational, which we denote as: $\mathcal{G} := \mathcal{E} \times \mathcal{A} \times \mathcal{P} \times \mathcal{E} \times \mathcal{A} \times \mathcal{T}$, where $\mathcal{E}, \mathcal{P}, \mathcal{A}$ and $\mathcal{T}$ denote the entity, predicate, attribute, and timestamp set.

When we consider an *event* occurring in a video, it can be regarded as a tuple that describes a holistic fact. We use a tuple to denote a single event e:

$$e = \cup(s, sa, p, o, oa, ts) \quad s, o \in \mathcal{E}, p \in \mathcal{P}, sa, oa \in \mathcal{A}, ts \in \mathcal{T} \quad (1)$$

For video-language models, an event e is described in natural language. In this paper, for the sake of clarity in writing, *event* can refer to both event tuples and textual captions that describe an event within the given context. We denote event tuples as e and textual descriptions as $t(e)$ to distinguish between the two.

Measuring Alignment Representational alignment is an active research topic in multi-modal learning. For video-language models, we pay more attention to the cross-modality alignment that leads to shared semantic information between video and texts.

To measure the video and text representational alignment in the semantic space, we use the similarity-based metrics of video and text embeddings as a common approach(Sucholutsky et al., 2023; Radford et al., 2021).

Video-text Alignment at Event Level As videos are dynamic graphs with multiple entities and semantic relations, aligning videos v and any $e \subseteq \mathcal{G}$ is important. Considering the properties of video data, we propose the following three aspects of video-text alignment at event level:

Temporal alignment: temporal ordering is vital in a video's semantic meaning; this can involve the temporal ordering of predicates (the sequence of two actions) or attributes (the change of entity status) in an event;

Neighborhood alignment: videos usually involve multiple entities and relations. Distinguishing confounding entities and their attribution binding is important for video-language alignment. This can also be regarded as "spatial" alignment between video and texts;

Factual alignment: videos also contain highly semantic information in their entity/predicate/attributes, which means inconsistent texts that contradict the facts in videos should be seen as misalignment.

2.2 Metrics

To quantify the alignment of video and events, we propose using an **alignment score** defined as:

$$score = \frac{\sum_i^N I(s(\mathbf{v}_i, \mathbf{t}(e_i^+)) > s(\mathbf{v}_i, \mathbf{t}(e_i^-)))}{N} \quad (2)$$

$s(\cdot)$ is a similarity measure between video features $\mathbf{v}$ and text features $\mathbf{t}$, *e.g*cosine similarity. We denote a truthful event occurring in a video v_i as e_i^+ and an inconsistent event as e_i^-, which is a *manipulated* event based on e_i^+. To bridge the gap between event tuples and natural language, we employ a decorator $t(\cdot)$ to generate natural sentences, which will be introduced in Sec. 3.1.

We determine the alignment score by the ratio of alignment to misalignment in a set of videos where video-language models correctly apply a higher

semantic similarity to truthful events than manipulated ones. An alignment score of 1 demonstrates that video-language models can accurately align videos and texts, whereas an alignment score of 0 indicates that video-language models struggle to differentiate between truthful and manipulated events, resulting in poor alignment. We aim to calculate a score for each alignment type to assess event-level alignment in different aspects.

2.3 Event Manipulation

We need to manipulate truthful events e^+ as counterexamples e^- to measure the alignment scores. To measure the above-mentioned event-level alignment type respectively, we define three manipulation types as illustrated in Fig. 2:

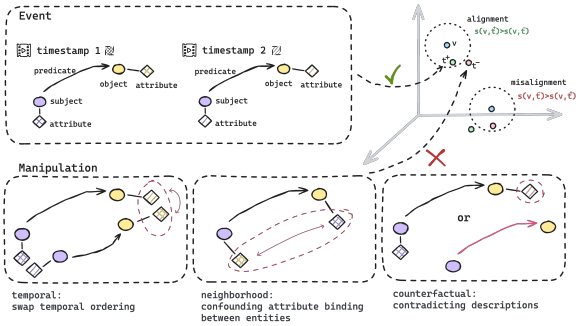

Figure 2: **We define three manipulation types**: temporal, neighborhood, and counterfactual manipulations are targeted at the three alignment types we are interested in. By manipulating event tuples, we generate an inconsistent event e^- with the video content, which is supposed to be dissimilar to the video representation v in the semantic space. Misalignment is identified if video-language models have higher similarity between video and manipulated events than truthful event: $s(\mathbf{v}, \mathbf{t}^-) > s(\mathbf{v}, \mathbf{t}^+)$.

Temporal Manipulation is devised to *swap* temporal ordering of two elements like predicates and attributes. This highlights whether the video-language models can sense the temporal ordering of two events, a key aspect of videos.

For two events (s^1, p^1, o^1, ts^1) and (s^2, p^2, o^2, ts^2) in the scene graph, the temporal manipulation $e^{-\text{predicate}}_{\text{temporal}}$ and $e^{-\text{attribute}}_{\text{temporal}}$ are defined as

$$\begin{aligned} e^{-\text{predicate}}_{\text{temporal}} &= \texttt{Swap}\left((s^1, p^1, o^1, ts^1) \cup (s^2, p^2, o^2, ts^2)\right) \\ &= (s^1, p^1, o^1, ts^2) \cup (s^2, p^2, o^2, ts^1) \end{aligned} \tag{3}$$

$$\begin{aligned} e^{-\text{attribute}}_{\text{temporal}} &= \texttt{Swap}\left((s^1, sa^1, ts^1) \cup (s^1, sa^2, ts^2)\right) \\ &= (s^1, sa^2, ts^1) \cup (s^1, sa^1, ts^2) \end{aligned} \tag{4}$$

Neighborhood Manipulation is crafted to *swap* attributes among distinct neighboring entities in the video. One scene in the video usually involves multiple entities, like different characters or objects. This highlights the ability to align attributes correctly in videos. For an event (s^1, sa^1, o^1, oa^1) in the scene graph, the neighborhood manipulation $e^{-\text{attribute}}_{\text{neighborhood}}$ is defined as

$$\begin{aligned} e^{-\text{attribute}}_{\text{neighborhood}} &= \texttt{Swap}\left((s^1, sa^1, o^1, oa^1)\right) \\ &= (s^1, oa^1, o^1, sa^1) \end{aligned} \tag{5}$$

Counterfactual Manipulation is targeted at replacing an event element with a *counterfactual* concept that contradicts the truth. For an event (s^1, sa^1, p^1, o^1) in the scene graph, the counterfactual manipulation $e^{-\text{predicate}}_{\text{counterfactual}}$ and $e^{-\text{attribute}}_{\text{counterfactual}}$ are defined as

$$\begin{aligned} e^{-\text{predicate}}_{\text{counterfactual}} &= \texttt{Counter}\left((s^1, sa^1, p^1, o^1)\right) \\ &= (s^1, sa^1, p^1_{\text{-}}, o^1) \end{aligned} \tag{6}$$

$$\begin{aligned} e^{-\text{attribute}}_{\text{counterfactual}} &= \texttt{Counter}\left((s^1, sa^1, p^1, o^1)\right) \\ &= (s^1, sa^1_{\text{-}}, p^1, o^1) \end{aligned} \tag{7}$$

where $sa^1_{\text{-}}$ and $p^1_{\text{-}}$ represent generated counterfactual attributes and predicates that contradict the fact.

3 Experiments

3.1 Datasets

In view of the demand for rich scene graph annotations to extract events, we consider two widely used video-text datasets in video understanding research: ActivityNet(Caba Heilbron et al., 2015) and MovieGraphs(Vicol et al., 2018). To extract event tuples, we utilize additional video scene graph annotations in (Yu et al., 2023; Krishna et al., 2017) to augment the data sources. Datasets statistics are shown in the Tab. 4.

Manipulation Details To control the consistency of the manipulated and original events, we only manipulate the predicates/attributes of the same fine-grained type as in Tab. 1-2. We follow the categorization of predicates and attributes in the original datasets' annotation for consistency. For predicates, we particularly generate subsets for each subcategory, which pay attention to the research focus of action understanding in videos. "Interaction" means predicates that involve more than one entity (subject and object), whereas "Action" only

involves one entity (subject); relationship mainly refers to the spatial or interpersonal relationship like family, friendship, and romance. Considering the three manipulation methods defined and the fine-grained predicate types, we identify **a total of 8 manipulation types**, as listed in Tab. 1-2.

Dataset	Predicate	Attribute
ActivityNet	Interaction	Occupation
		Gender
	Action	Age
	Spatial Relationship	Color
		Material

Table 1: Fine-grained types of predicates and attributes of ActivityNet.

Dataset	Predicate	Attribute
MovieGraphs	Interaction	Occupation
		Gender
	Action	Age
		Ethnicity
	Relationship	Emotion
		Appearance

Table 2: Fine-grained types of predicates and attributes of MovieGraph.

LLM Decorator Event tuples are still far from natural language for video-language models. Given a tuple, we employ GPT-3.5-turbo[1] as a controlled decorator $t(\cdot)$ to convert tuples into natural language. We first generate a sentence from tuples based on rules and then use the prompt "*rewrite the following sentence and make it more natural and grammatically correct, but you cannot add and alter any new information*" and setting the temperature coefficient as 1 to improve the sentence fluency of natural language. Several examples are delineated in the Appendix. For counterfactual manipulation, we also employ GPT-3.5-turbo to generate contradicting verbs/attributes/spatial prepositions with prompts as "*In this task, you are given a sentence, your job is to replace the verb/adjective/preposition in the sentence with a verb/adjective/preposition which still makes this sentence make sense.*".

3.2 Model Selection

We evaluate 6 video-language models: two variants of CLIP4Clip, Singularity, X-CLIP, ALPRO, and InternVideo, which are detailedly introduced in Sec. A.2. The first five are fine-tuned on our datasets for domain adaptation to ensure the model achieves comparable performance. InternVideo is a large-scale pre-trained video-language model evaluated in zero-shot settings.

3.3 Analysis

We report the alignment scores of all six model types in Fig. 3-4. A green background denotes a high alignment score and red denotes a low score.

[1]https://platform.openai.com/docs/models/gpt-3-5-turbo

3.3.1 Anaslysis on Event-level Alignment

We first take a deeper look at how video-language models perform on different event-level alignments.

All models show a poor temporal alignment, especially for predicates, including interactions and actions. Particularly, comparing alignment scores between temporal and counterfactual interaction/action manipulation, we speculate that video-language models have a fair action understanding ability but fail to distinguish the temporal ordering of actions. We conclude that temporal alignment is a vital but challenging aspect for video-language models; this also aligns with prior studies(Bagad et al., 2023), which shows that visual manipulations such as shuffling video frames also lead to video-text misalignments.

Similarly, neighborhood alignment turns out to be a challenge to most of the models. It can be viewed as a type of "spatial" alignment, requiring video-language models to acquire entity-level knowledge for each frame. Unfortunately, most models demonstrate a subpar alignment capability in this regard across both datasets. The challenge is further amplified in MovieGraph, as movie scenes frequently feature multiple characters, making differentiation more difficult.

Factual alignment shows comparatively good results. This indicates that video-language models can learn semantic meanings well with the BERT textual encoder. However, larger models' performance, like InternVideo cannot scale up with training on a larger corpus of data; this indicates larger foundation models are not perfect in the current stage for factual alignment. Further, we also find that video-language models cannot distinguish spatial relationships between objects in a video as shown in Tab. 3. This is further proof that video-language models have a limited ability of "spatial" alignment to understand frame-level information.

3.3.2 Comparison of Models

We analyze models from visual and temporal modeling (video encoders) and semantic modeling (text encoders). Among all models, the poor temporal and neighborhood alignments show that **video encoders are not powerful enough to represent videos**. This involves two-fold: event-ordering understanding for temporal alignment and entity-level understanding for neighborhood alignment. We attribute this to the sparse frame sampling method: sparse sampling might work on other tasks like

Video Retrieval and some existing Video Question Answering benchmarks(Lei et al., 2022; Luo et al., 2022); as a result, static video content achieves high results on these benchmarks. However, for event-level alignment, naive sparse sampling is inadequate. Furthermore, temporal and neighborhood alignments highly depend on the visual encoder for single frames. Most models use a pre-trained CLIP encoder; however, since prior works(Ma et al., 2022a; Goel et al., 2022) in the image domain suggest that CLIP still lacks entity-level knowledge, our results show the problem is also not negligible for video-language alignment. For word-level knowledge for factual alignment, all models use BERT as a text encoder: we see that video-language models can learn confounding semantic meanings and achieve especially good factual alignment in general when trained with large-scale data.

In particular, large video foundation models like InternVideo also suffer from poor temporal and neighborhood alignment and its factual alignment ability is not prominent in comparison with other fine-tuned models. The training scale of 12 million video clips in 5 domains (Wang et al., 2023c) does not promise an event-level alignment for video-language models. We speculate that data diversity is as important as data scale in training video-language models.

3.4 Discussion

In general, we conclude that **temporal alignment and neighborhood alignment are bad** for video-language models while training on video-language datasets can effectively help video-language models understand word-level knowledge for predicates and verbs. We attribute these issues to two reasons: (1) video-language models lack an effective visual encoder to understand entity-level and temporal semantics; (2) training datasets lack diversity for video-language models to learn misalignments.

4 Improving Event-level Video-Language Alignment

Poor event-level video-language alignment reveals that video-language models do not learn a perfect representation, particularly at a fine-grained level. Considering that video-language models are pre-trained in a manner that semantically aligned samples should be encouraged to have higher similarity and misaligned samples should be distinguished, we wonder if plugging manipulated events as misaligned samples can help improve the event-level alignment, as in (Momeni et al., 2023; Radenovic et al., 2023; Huang et al., 2023).

In this section, we propose leveraging misaligned events as hard negative training to see if these hard negative samples can effectively help video-language models.

4.1 Experiments

We utilize a two-stage pipeline to post-train the video-language model with hard negative samples. (1) we post-train the video and text encoders using video-text contrastive learning with hard negative samples; (2) we freeze the encoders and incorporate task-specific decoders for downstream tasks.

Prototype Model We selected a two-stream structure as our base model, which consists of a ViT-based video encoder and a BERT-based text encoder. The encoders enhanced with post-training using hard negative samples are kept frozen while an additional transformer-based decoder is trained for the QA task.

Hard-Negative Video-Event Contrastive Loss Video-text Contrastive Learning is the predominant methodology for multi-modal alignment, as shown in all the models we evaluated. Drawing inspiration from (Momeni et al., 2023; Radenovic et al., 2023; Bagad et al., 2023; Chen et al., 2020), we (1) extract and manipulate events from the training set of the dataset in the same manner as in our dissecting phase and (2) employ the hard negative noise contrastive loss for event-level video-language alignment and perform a reweighting of negative samples according to their hardness. For each sample, we randomly sample $\mathcal{N}^i_{\text{gen}}$ misaligned events for e_i from all possible manipulation types.

Our new video-to-text alignment loss enhanced with generated hard negative samples is defined as:

$$\mathcal{L}_{v2t} = -\log\left(\exp s(v_i, t_i)\right) + \log\Bigg(\exp s(v_i, t_i) + \sum_{j\neq i}^{\mathcal{N}_{\text{in}}} \exp s(v_i, t_j)\cdot\omega_{i,j} + \sum_{k=1}^{\mathcal{N}_{\text{gen}}} \exp s(v_i, t_k)\cdot\omega_{i,k}\Bigg) \quad (8)$$

here $\mathcal{N}_{\text{in}}$ denotes the set of in-batch negative samples, and $\mathcal{N}_{\text{gen}}$ is the set of generated misaligned events as hard negative samples based on t_i. We also reweight the negative samples according to their hardness as in (Radenovic et al., 2023) and preserve the original text-to-video alignment loss:

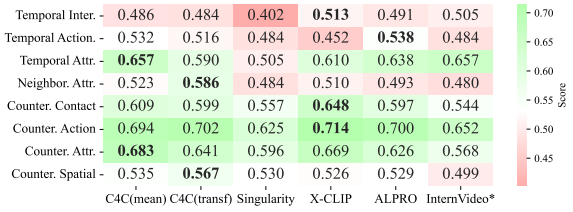

Figure 3: **Alignment score on ActivityNet.** We highlight the alignment scores higher than 0.6 in green and lower than 0.52 in red. For each manipulation type, the highest score is in **boldface** in each row. *InternVideo shows zero-shot performance.

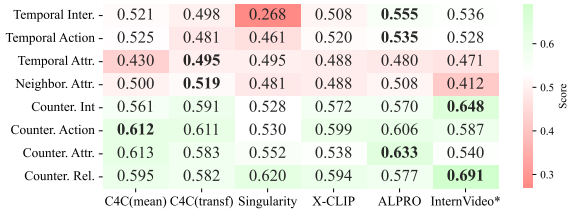

Figure 4: **Alignment Score on MovieGraph.** We highlight the alignment scores higher than 0.56 in green and lower than 0.5 in red. For each manipulation type, the highest score is in **boldface** in each row. *InternVideo shows zero-shot performance.

a detailed formulation can be found in the Supplementary Material.

Implementation Details We initialize the vision encoder with the $\text{BEiT}_{\text{BASE}}$(Bao et al., 2022) pre-trained on ImageNet-21K (Deng et al., 2009). Meanwhile, the text encoder is initialized with the first nine layers of $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019).

Training Details We first conduct a hard-negative post-training with the loss mentioned above. We use 10 epochs and AdamW (Loshchilov and Hutter, 2019) optimizer with an initial learning rate of 1e-4 for training. In the first epoch, a learning rate warm-up is applied, followed by cosine decay (Loshchilov and Hutter, 2017) to 1e-6 for the remainder of the training. We used a total batch size of 64; we sampled 7 generated negative samples for each sample. We also conduct task-specific fine-tuning to evaluate the post-trained encoders on a downstream QA task. For the Question-Answering (QA) task, we keep both encoders from the post-training frozen and only train the text decoder. We train for 10 epochs with AdamW with an initial learning rate of 1e-5 and warm up the learning rate in the first half epoch, followed by cosine decay to 1e-6. We use a batch size of 64. For both tasks, the model inputs 4 frames.

4.2 Experimental Results

We first show that after post-training, most event-level alignment types have been improved, as shown in Fig.5. However, we also observed that the post-training has a counter effect for temporal attribute and temporal action alignment. This aligns with our previous analysis: for temporal alignment, data-driven approaches are insufficient for video-language models to represent dynamic videos better; on the contrary, more efficient temporal modeling is vital.

To better inspect the model's event-level alignment ability after post-training, we evaluate the model on a new Video QA benchmark, AnetQA (Yu et al., 2023), which features answering fine-grained and temporal questions. As shown in Tab. 3, post-training is beneficial for the Video QA task. This indicates that better alignment can also help improve the downstream task.

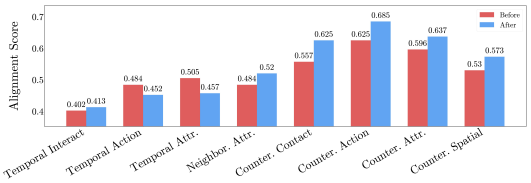

Figure 5: **Alignment scores before/after post-training with misaligned events as hard negative samples.** Neighborhood and factual alignment have been improved after post-training; however, post-training does not help with temporal alignment.

4.3 Discussion

Fig. 5 shows that post-training with negative samples as a data-driven method can improve event-level alignment in multiple aspects, including neighborhood and factual alignments, and yield better performance in the downstream tasks. However, it also implies again that video-language models lack good temporal alignment due to their visual encoder not being an effective model for temporal semantic information; simply, data-driven learning objectives cannot compensate for the importance of model architecture.

Limitations We showed that augmenting event-level negative samples is not a conclusive solution: on the one hand, the overhead in mining event-level video data annotation persists. Given the data scale in pre-training video-language models, the pivotal issue of efficiently providing clean

and diverse negative data for video-language pre-training remains an open question. On the other hand, data-driven methods still struggle to learn temporal-aware video representations.

	Accuracy (%)		
	temporal	counterfactual	overall
w/o Manipulated Events	29.18	3.80	21.47
w/ Manipulated Events	**33.63**	**5.55**	**22.34**

Table 3: **The results on the test subset of AnetQA.** We observe that better alignment with misaligned events for post-training also benefits downstream Video QA tasks, including temporal and counterfactual questions.

5 Related Works

Video-Language Model By learning a joint cross-modal representation, video-language models bridge the gap between video representation and natural language. Works have shown that with supervised video-text pairs (Wang et al., 2023c; Bain et al., 2021b; Zeng et al., 2023), video-language models can learn good joint representations for downstream tasks like Video-Text Retrieval (Luo et al., 2022; Lei et al., 2022; Bain et al., 2021b), Video Captioning (Sun et al., 2019) and Video Question Answering (Wang et al., 2023a; Dongxu Li, 2022; Lei et al., 2022). Recently, video foundation models (Wang et al., 2022) trained by larger-scale video-text data show excellent generalization performance on downstream tasks.

Understanding Vision-Language Model In recent years, along with the popularity of vision-language models, this has been an important topic in analyzing and probing their representation ability. For the image domain, Hendricks and Nematzadeh (2021); Yuksekgonul et al. (2022); Thrush et al. (2022); Li et al. (2022b); Ma et al. (2023); Zhao et al. (2022); Gandhi et al. (2022); Cascante-Bonilla et al. (2023); Li et al. (2023b); Nikolaus et al. (2022) design different benchmarks centered on the compositional representation between images and texts: Hendricks and Nematzadeh (2021); Li et al. (2022b); Ma et al. (2023); Zhao et al. (2022) aim to probe the components like relation and attribute understanding in an image-text pair, While Yuksekgonul et al. (2022) focuses on multiple linguistic perspectives on the sentence element ordering, Thrush et al. (2022); Cascante-Bonilla et al. (2023) further brings the semantic perturbation to images with image editing or synthesized images. Dynamic scenes and multiple events make videos more complicated and hard to probe. Momeni et al. (2023); Park et al. (2022) investigates if video-language models can tell the nuance difference between verbs. Bagad et al. (2023) probes if a video-language model is sensitive to temporal ordering. However, these pioneering works are on a *broad level* and have a single perspective.

Data-driven methods have been widely adopted in previous research (Ma et al., 2023; Bagad et al., 2023; Wang et al., 2023b; Chen et al., 2020; Yuksekgonul et al., 2022; Sahin et al., 2024; Bansal et al., 2023b; Doveh et al., 2023; Huang et al., 2024) with augmented negative mining and also manifest benefits on QA tasks (Chen et al., 2020; Gandhi et al., 2022). This also inspired us to improve event-level video-language alignment. But in our analysis, we should notice that data-driven methods do not solve the problem for video data once and for all.

Video Scene Graphs in Video-Language Dataset Video scene graphs represent videos' dynamic and multi-event nature in a graph format. However, annotating video scene graphs is time and labor-intensive work. Benefit from the contributions of the community, Wu et al. (2021); Yu et al. (2023); Vicol et al. (2018) provide dense and timestamped scene graph-level annotations with rich types of predicates and attributes on top of video data, which enables us to extract events and construct our benchmark.

6 Conclusion

In this work, we propose a new framework to dissect the event-level video-language alignment of 6 existing video-language models on two datasets into three aspects: temporal, neighborhood, and factual alignment. Our in-depth analysis reveals that existing contrastive methods for video-language pre-training have a poor temporal and neighborhood alignment and comparably better but non-perfect factual alignment. This means that video-language models still lack event-level knowledge to understand video's dynamic and complex structure. More importantly, we show that data-driven approaches with more fine-grained and rich annotation are essential for better video-language representation learning but do not compensate for the importance of better video temporal modeling innovation.

Limitations

This work relys on web video datasets consisting of online and movie videos, which represent a limited subset of available video data. Expanding to more diverse and complex video datasets in other domains, such as egocentric videos, autonomous driving, gaming videos, will be the focus of future work. Besides, a human baseline for evaluating such alignment is also of interests, which serves as the upper bound of event-level alignment.

References

Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. 2023. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2503–2516.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021a. 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*, pages 1728–1738.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021b. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. 2023a. Videocon: Robust video-language alignment via contrast captions. *Preprint*, arXiv:2311.10111.

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. 2023b. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. *arXiv preprint arXiv:2303.03323*.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. 2023. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10800–10809.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. 2009 ieee conference on computer vision and pattern recognition.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Hongdong Li Juan Carlos Niebles Steven C.H. Hoi Dongxu Li, Junnan Li. 2022. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2022. Measuring Compositional Consistency for Video Question Answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5036–5045, New Orleans, LA, USA. IEEE.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. 2023. Structure-CLIP: Enhance Multi-modal Language Representations with Structure Knowledge. *Preprint*, arxiv:2305.06152.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. 2024. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *AAAI*. AAAI Press.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and

vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. *Preprint*, arXiv:2206.03428.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022a. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022b. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. *Preprint*, arXiv:1608.03983.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022a. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18051–18061.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022b. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591.

Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. 2022. Do vision-and-language transformers learn grounded predicate-noun dependencies? *arXiv preprint arXiv:2210.12079*.

Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. 2022. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6967–6977.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. 2024. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint

model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8581–8590.

Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023a. All in one: Exploring unified video-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Weihan Wang, Zhen Yang, Bin Xu, Juanzi Li, and Yankui Sun. 2023b. Vilta: Enhancing vision-language pre-training through textual augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3158–3169.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. 2023c. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Bo Wu, Shoubin Yu, Tenenbaum Joshua B Chen, Zhenfang, and Chuang Gan. 2021. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*.

Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. 2023. Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23191–23200.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Ziyun Zeng, Yuying Ge, Xihui Liu, Bin Chen, Ping Luo, Shu-Tao Xia, and Yixiao Ge. 2023. Learning transferable spatiotemporal representations from natural script knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23079–23089.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations. *Preprint*, arxiv:2207.00221.

A Appendix

A.1 Datasets

In this section, we introduce the details of the dataset.

A.1.1 Statistics of Datasets

Dataset	Domain	Video Number	Average Duration	Total Captions	Average Caption Length
ActivityNet (Caba Heilbron et al., 2015)	Activity	11.5K	180s	16K	35.11
MovieGraphs (Vicol et al., 2018)	Movie	7.6K	44.28s	7.6K	8.1

Table 4: **Statistics of two Datasets.** We use two video-text datasets that are in different domains and vary in video and text distribution.

A.1.2 Statistics of Manipulation

According to the videos' content and annotation, we have different sizes of varied manipulation sets, which are listed in Tab. 5-6 for both datasets.

Manipulation Type	Number of Samples
temporal contact	184
temporal action	62
temporal attribute	102
neighborhood attribute	35
counterfactual contact	1168
counterfactual action	1008
counterfactual attribute	818
counterfactual spatial relationship	935

Table 5: Number of event tuples extracted from ActivityNet in each manipulation category in our benchmark.

Manipulation Type	Number of Samples
temporal interaction	563
temporal action	489
temporal attribute	160
neighborhood attribute	597
counterfactual relationship	583
counterfactual interaction	591
counterfactual action	535
counterfactual attribute	591

Table 6: Number of event tuples extracted from MovieGraphs in each manipulation category in our benchmark.

A.1.3 Examples of MovieGraphs

In Tab. 7-8, we showcase some examples of manipulated sentences we generate in the ActivityNet and MovieGraphs datasets.

A.1.4 License

The dataset and code will be made accessible. We will provide the manipulated version of video captions for the purpose of scientific research.

We use standard licenses from the community and provide the following links to the non-commercial licenses for the datasets we used in this paper.

ActivityNet: https://github.com/activitynet/ActivityNet/blob/master/LICENSE

MovieGraph: http://moviegraphs.cs.toronto.edu/index.html

A.2 Model Comparison

A systematic model comparison on different dimensions is listed in Tab. 9. We compare selected models in terms of visual encoder, textual encoder, pre-training objectives, temporal modeling methods, input frames, and whether they are finetuned.

CLIP4Clip (Luo et al., 2022) is built on the CLIP framework and explores different temporal modeling modules to address video-text retrieval. Our experiments employ two distinct similarity calculators: a parameter-free type, specifically mean pooling and a sequential type based on the Transformer Encoder. For brevity, we denote them as C4C(mean) and C4C(transf).

Singularity (Lei et al., 2022) exploits the single-frame bias of videos and adopts an extremely sparse sampling method. We use the Singularity-temporal variant described in (Lei et al., 2022) with a 2-layer transformer-based temporal encoder; we use Singularity to denote this variant for brevity.

X-CLIP (Ma et al., 2022b) stands out as a multi-grained contrastive video-language model. It aims to aggregate fine-grained (video-sentence and frame-word) and cross-grained (video-word and frame sentence) alignment to improve video alignment.

ALPRO (Dongxu Li, 2022) uses sparsely sampled video frames and leverages entity-object level fine-grained details to align video language pairs.

InternVideo (Wang et al., 2023c) serves as a comprehensive video foundation model. It adopts a transformer-based video and text encoder trained on contrastive learning loss and masked vision loss. We use the official implementation and pre-trained model of InternVideo with a CLIP-based ViT as the video encoder.

A.2.1 Extended Model Evaluation

We report the full tables of alignment scores on ActivityNet and MovieGraphs in Tab. 10-11. We

Type	Positive	Negative
Temporal Contact	A person with white skin is kneeling on a yellow yard before eventually reaching out to touch the brown mulch	A person with white skin is kneeling in the yellow yard, while touching the brown mulch
Temporal Act.	A silver metal knife is sliding, and after a while, a white oven is opened	A silver metal knife is sliding as a white oven is being opened
Temporal Attr.	A yellow bike after a while turns black	A bike which is black, after a while its color is yellow
Neighbor. Attr.	The color of the smoke ring is white, and the pipe is brown	The smoke ring is brown in color, while the pipe is white in color
Counter. Contact	A white chalk is drawn on the white street	The white chalk is being erased from the white street
Counter. Act.	A green cloth punchbag is hanging	A green cloth-made punchbag is swinging
Counter. Attr.	A black bike is parked	A shiny bike is parked
Counter. Spatial.	A boy contestant is on a white hill.	A boy contestant is standing beside the white hill

Table 7: We showcase the positive and negative captions generated from extracted and manipulated event tuples on the ActivityNet dataset. We highlight the truthful information in cyan and manipulated information in magenta.

Type	Positive	Negative
Temporal Inter.	Mikael Blomkvist who is male journalist asks Lisbeth Salander, before Lisbeth Salander apologizes Mikael Blomkvist	Mikael Blomkvist who is male journalist asks Lisbeth Salander, after Lisbeth Salander apologizes Mikael Blomkvist
Temporal Act.	Workers cut branches off a tree, before Workers shred branches	Workers cut branches off a tree, during Workers shred branches
Temporal Attr.	Tommy is excited after a while Tommy becomes skeptical	Tommy is skeptical after a while Tommy gets excited
Neighbor. Attr.	A male Truck Driver is annoyed and the Security Guard looks responsible and calm	A male Truck Driver looks responsible and calm while the Security Guard looks annoyed.
Counter. Inter.	Joey Donner talks to Bianca Stratford, Joey Donner, who is male	Joey Donner flirts with Bianca Stratford, Joey Donner who is male
Counter. Act.	Greg Focker carries lawn chairs	Greg Focker assembles lawn chairs
Counter. Attr.	Cecilia is honest woman	Cecilia is irresponsible woman
Counter. Relation.	Lisbeth Salander is colleague of Mikael Blomkvist	Lisbeth Salander is the teacher of Mikael Blomkvist

Table 8: We showcase the positive and negative captions generated from extracted and manipulated event tuples on the MovieGraphs dataset. We highlight the truthful information in cyan and manipulated information in magenta.

Model	**Visual Encoder**	**Textual Encoder**	**Pre-Training Objectives**	**Temporal Modeling**	**Input Frames**	**Fine-tuned Modeling**
C4C(mean)	ViT-B/32	BERT	VTC	Mean-Pooling	12	✓
C4C(transf)	ViT-B/32	BERT	VTC	Temporal transformer	12	✓
Singularity	$\text{BEiT}_{\text{BASE}}$	BERT	VTC+VTM +MLM	Temporal transformer (2-layer)	4	✓
X-CLIP	ViT-B/32	BERT	VTC	Temporal transformer	12	✓
ALPRO	ViT-B/16	$\text{BERT}_{\text{BASE}}$	VTC+VTM +MLM+PEM	TimeSformer(Bertasius et al., 2021) +Mean-Pooling	8	✓
InternVideo	ViT-L/14	BERT	VTC	Linear Projection	8	✗

Table 9: **Comparison of model architecture.** We compare five selected models (6 variants) on different dimensions: visual encoder, textual encoder, pre-training objectives, temporal modeling, and if we fine-tune the model on the datasets.

Model	C4C(mean)	C4C(transf)	Singularity	X-CLIP	ALPRO	InternVideo*
Temporal Inter.	0.486(±0.08)	0.484(±0.02)	0.402(±0.01)	**0.513**(±0.05)	0.491(±0.04)	0.505(-)
Temporal Action.	0.532(±0.04)	0.516(±0.02)	0.484(±0.02)	0.452(±0.03)	**0.538**(±0.04)	0.484(-)
Temporal Attr.	**0.657**(±0.00)	0.590(±0.01)	0.505(±0.08)	0.610(±0.02)	0.638(±0.07)	**0.657**(-)
Neighbor. Attr.	0.523(±0.06)	**0.586**(±0.02)	0.484(±0.01)	0.510(±0.02)	0.493(±0.03)	0.480(-)
Counter. Contact	0.609(±0.02)	0.599(±0.02)	0.557(±0.01)	**0.648**(±0.03)	0.597(±0.02)	0.544(-)
Counter. Action	0.694(±0.00)	0.702(±0.02)	0.625(±0.01)	**0.714**(±0.02)	0.700(±0.02)	0.652(-)
Counter. Attr.	**0.683**(±0.01)	0.641(±0.02)	0.596(±0.01)	0.669(±0.05)	0.626(±0.00)	0.568(-)
Counter. Spatial	0.535(±0.03)	**0.567**(±0.03)	0.530(±0.01)	0.526(±0.01)	0.529(±0.01)	0.499(-)

Table 10: **Alignment score on ActivityNet.** We highlight the alignment scores higher than 0.6 in green and lower than 0.52 in red. For each manipulation type, the highest score is in **boldface** in each row. *InternVideo shows zero-shot performance.

Model	C4C(mean)	C4C(transf)	Singularity	X-CLIP	ALPRO	InternVideo*
Temporal Inter.	0.521(±0.01)	0.498(±0.00)	0.268(±0.00)	0.508(±0.02)	**0.555**(±0.02)	0.536(-)
Temporal Action	0.525(±0.02)	0.481(±0.01)	0.461(±0.02)	0.520(±0.02)	**0.535**(±0.01)	0.528(-)
Temporal Attr.	0.430(±0.01)	**0.495**(±0.02)	**0.495**(±0.02)	0.488(±0.01)	0.480(±0.01)	0.471(-)
Neighbor. Attr.	0.500(±0.01)	**0.519**(±0.01)	0.481(±0.03)	0.488(±0.06)	0.508(±0.02)	0.412(-)
Counter. Int	0.561(±0.01)	0.591(±0.01)	0.528(±0.01)	0.572(±0.01)	0.570(±0.01)	**0.648**(-)
Counter. Action	**0.612**(±0.00)	0.611(±0.01)	0.530(±0.01)	0.599(±0.01)	0.606(±0.01)	0.587(-)
Counter. Attr.	0.613(±0.02)	0.583(±0.01)	0.552(±0.01)	0.538(±0.04)	**0.633**(±0.02)	0.540(-)
Counter. Rel.	0.595(±0.03)	0.582(±0.02)	0.620(±0.00)	0.594(±0.01)	0.577(±0.02)	**0.691**(-)

Table 11: **Alignment score on ActivityNet.** We highlight the alignment scores higher than 0.6 in green and lower than 0.52 in red. For each manipulation type, the highest score is in **boldface** in each row. *InternVideo shows zero-shot performance.

repeat each experiment with three random seeds and report the average performance and standard deviation in the table.

A.3 Hard-negative Post-training Loss

We have defined the hard-negative post-training loss in the main body as:

$$\begin{aligned}\mathcal{L}_{v2t} &= -\log\left(\exp s(v_i, t_i)\right) + \log\Bigg(\exp s(v_i \cdot t_i) \\ &\quad + \sum_{j\neq i}^{\mathcal{N}_{\text{in}}} \exp s(v_i, t_j)\cdot \omega_{i,j}^{v2t} + \sum_{k=1}^{\mathcal{N}_{\text{gen}}} \exp s(v_i, t_k)\cdot \omega_{i,k}^{v2t}\Bigg), \\ \mathcal{L}_{t2v} &= -\log\left(\exp s(t_i, v_i)\right) + \log\Bigg(\exp s(t_i \cdot v_j) \\ &\quad + \sum_{j\neq i}^{\mathcal{N}_{\text{in}}} \exp s(t_i, v_j)\cdot \omega_{i,j}^{t2v}\Bigg),\end{aligned} \tag{9}$$

Inspired by Radenovic et al. (2023), we employ weight functions $\omega_{i,j}^{v\to t}$ and $\omega_{j,i}^{t\to v}$ formulated as:

$$\begin{aligned}\omega_{i,j}^{v2t} &= \frac{(\mathcal{N}_{\text{in}} + \mathcal{N}_{\text{gen}}^{i} - 1)\exp\left(\beta v^i \cdot t^j/\tau\right)}{\sum_{m\neq i}^{\mathcal{N}_{\text{in}}+\mathcal{N}_{\text{gen}}^{i}} \exp\left(v^i \cdot t^m/\tau\right)}, \\ \omega_{j,i}^{t2v} &= \frac{(\mathcal{N}_{\text{in}} - 1)\exp\left(\beta v^j \cdot t^i/\tau\right)}{\sum_{m\neq i}^{\mathcal{N}_{\text{in}}} \exp\left(v^m \cdot t^i/\tau\right)}\end{aligned} \tag{10}$$