Can Text-to-Video Generation help Video-Language Alignment?

Luca Zanella[1] Massimiliano Mancini[1] Willi Menapace[2]
Sergey Tulyakov[2] Yiming Wang[3] Elisa Ricci[1,3]
[1]University of Trento [2] Snap Inc. [3]Fondazione Bruno Kessler

https://lucazanella.github.io/synvita/

Abstract

Recent video-language alignment models are trained on sets of videos, each with an associated positive caption and a negative caption generated by large language models. A problem with this procedure is that negative captions may introduce linguistic biases, i.e.*, concepts are seen only as negatives and never associated with a video. While a solution would be to collect videos for the negative captions, existing databases lack the fine-grained variations needed to cover all possible negatives. In this work, we study whether synthetic videos can help to overcome this issue. Our preliminary analysis with multiple generators shows that, while promising on some tasks, synthetic videos harm the performance of the model on others. We hypothesize this issue is linked to noise (semantic and visual) in the generated videos and develop a method,* SYNVITA*, that accounts for those.* SYNVITA *dynamically weights the contribution of each synthetic video based on how similar its target caption is w.r.t. the real counterpart. Moreover, a semantic consistency loss makes the model focus on fine-grained differences across captions, rather than differences in video appearance. Experiments show that, on average,* SYNVITA *improves over existing methods on VideoCon test sets and SSv2-Temporal, SSv2-Events, and ATP-Hard benchmarks, being a first promising step for using synthetic videos when learning video-language models.*

1. Introduction

Video-language alignment (VLA) aims to model the relationship between video content and natural language descriptions [55], a fundamental multimodal task that enables various applications, such as video captioning [15] and video-text retrieval [47]. This task is challenging because it requires the models to recognize not only the entities but also their spatial and temporal relationships.

Recent approaches exploit multimodal large language models (MLLMs) to address VLA [3, 28, 29] by tasking the MLLM to answer whether a given video and description are aligned. While effective, such MLLMs often lack sufficient understanding of temporal dynamics, such as action types or temporal orders [10, 31]. This limitation also stems from the video-and-language datasets used for the MLLM pre-training, as they are biased towards frame-level semantics: the appearance of a single frame is often sufficient to infer the alignment with the textual caption [25, 35].

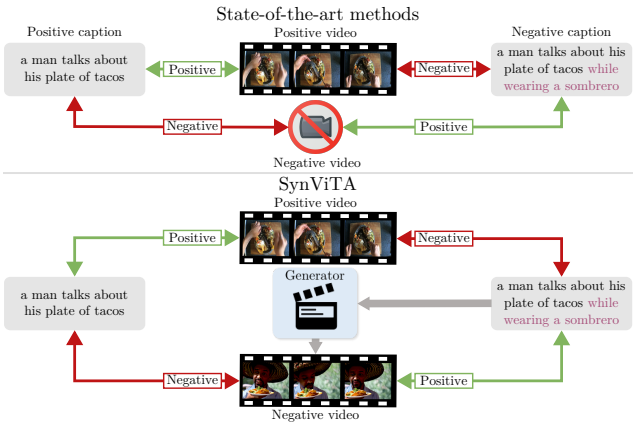

Figure 1. We study the problem of video-language alignment, *i.e.*, modeling the relationship between video content and text descriptions. Top: current methods use LLM-generated negative captions, which may introduce certain concepts (*e.g.*, *wearing a sombrero*) only as negatives, as they are not associated with any video. Bottom: we study whether overcoming this issue by pairing negative captions with generated videos can improve VLA.

While a possible solution is to augment datasets with negative captions, *e.g.*, captions from other videos, these negatives can be "easy" for VLA models to distinguish simply by focusing on nouns; therefore, recent works focus on LLM-generated captions as hard negatives [3]. However, relying solely on textual negatives may cause the MLLM to encounter concepts only as negatives, thus developing incorrect linguistic biases. For instance, in the VideoCon dataset [3], words like *sombrero*, *marshmallow*, and *bland* appear as textual negatives but not as positives. While

a remedy is to augment the training set with videos corresponding to the hard negative captions, retrieving such videos from existing databases is not a feasible solution as they lack sufficient videos that vary only w.r.t. actions or temporal order while remaining similar in all the other semantic aspects [35]. An alternative pathway is generating synthetic videos by feeding hard negative captions to text-to-video generative models [7, 34, 49, 58]. While this idea has been investigated in the image domain [36], it remains largely unexplored for videos.

In this paper, we aim to fill this gap and investigate, for the first time, the use of synthetic videos to improve VLA in temporal understanding. Specifically, we propose to leverage negative captions generated by existing models [3] and recent open-source text-to-video generators [7, 49, 58] to produce the corresponding synthetic videos (see Fig. 1). We first conduct a preliminary study to evaluate whether these generated videos can augment the training set of real videos and enhance performance on various video-related tasks. Our analysis shows that, while adding synthetic videos shows some promise, *it does not* consistently improve performance on temporally challenging downstream tasks, regardless of the generator. We also analyze the effects of different misalignment types (*i.e.*, semantically plausible changes in the video captions) on the generated videos. We notice that videos generated by, *e.g.*, introducing hallucination into the captions or reversing event order, align more with positive captions than with their target captions. Such noisy supervision signals may lead to ineffective learning, limiting improvements on downstream tasks.

Motivated by these preliminary findings, we argue that, when using synthetic videos for VLA we should account for (i) potential semantic inconsistency between input text and the generated videos and (ii) appearance biases, as synthetic videos may contain artifacts. We design SYNTHETIC VIDEOS FOR VIDEO-TEXT ALIGNMENT (SYNVITA), a model-agnostic method that can effectively tackle both challenges. SYNVITA addresses the semantic inconsistency problem by making the contribution of each synthetic video in the training objective proportional to their video-text alignment estimates [29]. Moreover, it accounts for appearance biases via a semantic regularization objective that (i) takes the common parts between the original and negative caption; (ii) encourages the model to focus on semantic changes rather than on the visual appearance difference between synthetic and real videos. We evaluate SYNVITA on the VideoCon [3] test sets with different Video LLMs [27, 60], and on temporally challenging downstream tasks, *i.e.*, text-to-video retrieval on SSv2-Temporal [42] and SSv2-Events [2] and video question answering on ATP-Hard [5]. On average, SYNVITA improves over state-of-the-art methods that do not use synthetic videos, demonstrating that synthetic videos can help VLA.

Contributions. To summarize, our contributions are:

- We pioneer the research problem in how to effectively leverage synthetic videos for VLA learning to improve temporal understanding;
- We conduct extensive analysis, shedding light on the potential benefits and limitations of using videos generated by state-of-the-art text-to-video generative models;
- We propose a new learning method for VLA with synthetic videos, SYNVITA, with a sample weighting strategy to mitigate noisy generations and a regularization term to enforce semantic understanding, instead of visual differences between synthetic and real videos.
- We evaluate SYNVITA on different benchmarks with different Video LLMs, proving its model-agnostic effectiveness in aiding VLA for better temporal understanding.

2. Related work

Video-language models for video understanding. Recent approaches for video understanding exploit the capabilities of foundation models. For instance, several works adapted CLIP [39], a model trained to compare images and texts, for video-language tasks, such as retrieval [14], captioning [33] or anomaly detection [62]. Other studies leveraged LLMs for reasoning over video captions [50, 63] or directly decode video features in natural language [27, 56, 64]. While these models heavily rely on pre-training on large-scale video-text pairs [55, 56], they still lack robustness in modeling temporal dynamics [10, 31]. Previous works addressed this by, *e.g.*, using LLMs to generate hard negatives [35], reversing the action sequence [2], or finer-grained objectives [51].

The closest work to ours is VideoCon [3], which finetunes a video LLM using temporally challenging hard *textual* negatives. However, our focus is different, as we explore whether generated videos can improve video-text alignment, complementing negative captions.

Video-language alignment evaluation. A main challenge in VLA is quantifying the semantic alignment between text and video frames. Early attempts used metrics based on the CLIPScore [19, 41, 43], which computes video-text alignment by measuring the similarity between video frames and their captions in the CLIP embedding space [39]. However, as VLMs struggle with temporal changes in captions [2, 35, 51, 61], recent approaches have started measuring video-text alignment using MLLMs for video question answering [3, 26, 28, 52, 53], such as the VQAScore in [29].

In this work, we use these models to evaluate the quality of the alignment and for the new objective of evaluating how much a synthetic video aligns with its textual counterpart.

Using synthetic visual data as training data. Recent works showed how augmenting training sets with synthetically generated images can improve the performance of discriminative models [17, 37, 45, 65]. Diffusion models,

known for their ability to generate highly realistic images and for their flexibility in dealing with different conditioning signals (text, depth, etc.), have significantly fostered this research trend [9]. While most works focused on image recognition tasks [36, 45, 65], recent approaches explored more challenging tasks such as few-shot recognition [17, 40] or out-of-distribution detection [13].

Our work follows a similar underlying idea and it is motivated by recent advances in text-to-video generation [7, 34, 49, 58]. However, we are the first to explore synthetic videos for improving video understanding models.

3. Video-language alignment

Video-language alignment aims to rate how well the content of a video matches a given text in natural language. Formally, let us define t as the given textual input in the language space $\mathcal{T}$, and V as a video in the space $\mathcal{V}$. The goal is to learn a function f parameterized by θ, mapping videos and texts to their alignment scores, *i.e.*, $f : \mathcal{V} \times \mathcal{T} \rightarrow [0, 1]$, where 1 means high alignment and 0 the opposite.

Given the fine-grained nature of language, this task requires video-language models with compositional and temporal order understanding and recent approaches use video LLMs for this task, where an LLM is used as decoder [3, 29]. Formally, let us define an LLM-based video-language model f via three functions: the visual encoder $f_{\texttt{vid}}$, the text encoder $f_{\texttt{txt}}$, and a decoder $f_{\texttt{dec}}$. The two encoders map their respective inputs into a shared d-dimensional embedding space, *i.e.*, $f_{\texttt{vid}} : \mathcal{V} \rightarrow \mathbb{R}^d$ and $f_{\texttt{txt}} : \mathcal{T} \rightarrow \mathbb{R}^d$. The decoder maps the visual and textual inputs into a vector in the probability simplex $\Delta^{|\mathcal{W}|}$ defined over the LLM vocabulary[1] $\mathcal{W}$, *i.e.*, $f_{\texttt{dec}} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \Delta^{|\mathcal{W}|}$. This probability vector is then used to sample the next token for the generative process.

Within this formulation, the alignment task becomes the probability of predicting `Yes` or `No` as the next word after the question π_q = `Does this video entail the description [t]?`, where $[t]$ is the target caption. Formally, this translates as f being:

$$f(V, t) = \frac{P_{\mathcal{W}}(\texttt{Yes}|V, t)}{P_{\mathcal{W}}(\texttt{Yes}|V, t) + P_{\mathcal{W}}(\texttt{No}|V, t)} \tag{1}$$

where $P_{\mathcal{W}}(\mathbf{w}|V, t) = f_{\texttt{dec}}^{[\mathbf{w}]}\left(f_{\texttt{vid}}(V), f_{\texttt{txt}}(\pi_q \circ t)\right)$, with π_q the shared question, $\circ$ string concatenation, and $f_{\texttt{dec}}^{[\mathbf{w}]}$ the likelihood of the word $\mathbf{w} \in \mathcal{W}$ from the decoder's output.

VLA learning. Usually, the parameters θ of f are updated using a dataset D of n video-language triplets $D = \{(V_1, t_1^+, t_1^-), \cdots, (V_n, t_n^+, t_n^-)\}$, where t_i^+ and t_i^- are the positive and negative text captions for the video V_i, *i.e.*, captions that respectively represent (t_i^+) and do not represent (t_i^-) the video content. Exploiting the probability distribution, output of $f_{\texttt{dec}}$, we can define the following objective:

$$\mathcal{L}_{\texttt{real}} = -\sum_{i=1}^{n} \log f(V_i, t_i^+) + \log\left(1 - f(V_i, t_i^-)\right). \tag{2}$$

This loss function forces f to sample `Yes` with a higher probability if the text represents the video and `No` otherwise.

4. Can synthetic videos help VLA?

The main loss function in VLA learning, as expressed in Eq. (2), considers as negative only textual inputs for a given "anchor" video. For each positive caption t_i^+, there is no negative video example associated with t_i^-. As such, linguistic biases might be induced in the MLLM because some concepts appear only as textual negatives. Thus, we wonder: *Can generated videos of negative captions help learning a VLA function?* To answer this question, we consider different text-to-video generator models and use them to generate synthetic videos associated to negative captions.

VLA learning with generated videos. Formally, a text-to-video generator G maps natural language expressions in $\mathcal{T}$ and noise in the space $\mathcal{N}$ to videos, *i.e.*, $G : \mathcal{T} \times \mathcal{N} \rightarrow \mathcal{V}$. For simplicity, we define $t^r = t^+$ (*i.e.*, text positively associated with the *real* video) and $t^s = t^-$ (*i.e.*, negative text for the real video, positively associated with the *synthetic* one). We propose to use the generator to define an objective over the dataset D:

$$\mathcal{L}_{\texttt{syn}} = -\sum_{i=1}^{n} \log f\left(V_i^s, t_i^s\right) + \log\left(1 - f\left(V_i^s, t_i^r\right)\right) \tag{3}$$

where $V_i^s = G(t_i^s, \eta_i)$, with $\eta_i \sim \mathcal{N}$ being the sampled noise. The negative text t_i^s is the input text to the generator, thus serving as the positive for the synthetic video, while the positive text t_i^r for the real video V_i^r serves as negative for the generated video.

Experimental analysis. To better understand the potential of synthetic videos, we first conduct a preliminary experimental analysis and leverage three state-of-the-art open-source video generators, *i.e.*, CogVideoX [58], LaVie [49], and VideoCrafter2 [7], to generate synthetic videos for each negative caption in the VideoCon dataset [3]. We augment the dataset with these generated videos and fine-tune a video LLM, mPLUG-Owl 7B [60], using the objective functions defined in Eq. (2) and Eq. (3) for real and synthetic videos, respectively. We measure the performance with the VLA scores estimated from Eq. (1), following the established evaluation protocol [3] across multiple tasks and datasets. Specifically, we consider video-language entailment on the VideoCon dataset, text-to-video retrieval on SSv2-Temporal [42] and SSv2-Events [2] datasets, and

[1] For simplicity, we omit the words' tokenization and we assume textual prompts and videos to be treated equally and encoded in the same space.

Text-to-Video Generator	Video-Language Entailment (VideoCon) LLM	Human	Human-Hard	Text-to-Video Retrieval SSv2-Temporal	SSv2-Events	Video QA ATP-Hard
None* [3]	**88.39**	77.16	74.76	13.00	10.37	35.46
CogVideoX [58]	83.93 (↓ 4.46)	76.89 (↓ 0.27)	75.10 (↑ 0.34)	11.76 (↓ 1.24)	8.79 (↓ 1.58)	35.30 (↓ 0.16)
LaVie [49]	85.26 (↓ 3.13)	76.96 (↓ 0.20)	74.63 (↓ 0.13)	**14.26** (↑ 1.26)	**10.80** (↑ 0.43)	34.82 (↓ 0.64)
VideoCrafter2 [7]	85.82 (↓ 2.57)	**77.33** (↑ 0.17)	**75.15** (↑ 0.39)	13.80 (↑ 0.80)	10.27 (↓ 0.10)	**35.79** (↑ 0.33)

Table 1. Results of the preliminary study on using synthetic videos generated by different text-to-video models. Increases (↑) and decreases (↓) are measured relative to the model fine-tuned without synthetic videos (*i.e.*, None). * indicates our reproduced results using the mPLUG-Owl 7B model checkpoint released in the original VideoCon repository.

Misalignment	Video-Language Entailment (VideoCon) LLM	Human	Human-Hard	Text-to-Video Retrieval SSv2-Temporal	SSv2-Events	Video QA ATP-Hard
None* [3]	**88.39**	77.16	74.76	13.00	10.37	35.46
Action	86.10 (↓ 2.29)	77.43 (↑ 0.27)	74.83 (↑ 0.07)	**15.04** (↑ 2.04)	10.66 (↑ 0.29)	36.28 (↑ 0.82)
Attribute	86.51 (↓ 1.88)	77.61 (↑ 0.45)	**75.50** (↑ 0.74)	13.67 (↑ 0.67)	11.47 (↑ 1.10)	35.25 (↓ 0.21)
Count	86.10 (↓ 2.29)	**77.66** (↑ 0.50)	75.27 (↑ 0.51)	14.27 (↑ 1.27)	10.97 (↑ 0.60)	36.16 (↑ 0.70)
Flip	85.69 (↓ 2.70)	76.04 (↓ 1.12)	73.53 (↓ 1.23)	14.94 (↑ 1.94)	10.73 (↑ 0.36)	36.06 (↑ 0.60)
Hallucination	85.46 (↓ 2.93)	76.55 (↓ 0.61)	74.77 (↑ 0.01)	13.89 (↑ 0.89)	10.14 (↓ 0.23)	**36.37** (↑ 0.91)
Object	**86.28** (↓ 2.11)	77.36 (↑ 0.20)	74.15 (↓ 0.61)	14.54 (↑ 1.54)	**11.54** (↑ 1.17)	35.48 (↑ 0.02)
Relation	86.22 (↓ 2.17)	77.46 (↑ 0.30)	74.59 (↓ 0.17)	14.99 (↑ 1.99)	11.38 (↑ 1.01)	34.65 (↓ 0.81)

Table 2. Average results of the preliminary study on using synthetic videos generated by different text-to-video models, for each type of misalignment. Increases (↑) and decreases (↓) are measured relative to the model fine-tuned without synthetic videos (*i.e.*, None). * indicates our reproduced results using the mPLUG-Owl 7B model checkpoint released in the original VideoCon repository.

video question answering (VQA) on the ATP-Hard dataset [5]. The evaluation metrics include the area under the receiver operating characteristic curve (AUC ROC) on video-language entailment, mean average precision (mAP) on text-to-video retrieval, and accuracy on VQA.

We report the results in Tab. 1, including baseline performance without synthetic video data (None). From the table, it is clear that synthetic videos harm the performance on the task closest to the training set (*i.e.*, average drop higher than 3% AUC on VideoCon LLM). One core reason for this drop is the distribution of the negatives being more similar to the one of the training set. Thus performance may decrease when a model sees them as positives. On the other hand, the results on downstream tasks suggest that synthetic videos hold promise. For instance, VideoCrafter2 improves the result of the baseline in 4/6 settings, while LaVie boosts performance on SSv2-Temporal (*i.e.*, +1.26 mAP). However, even with state-of-the-art video generators, not all of them guarantee improvements, and no single generator consistently outperforms the others across the tested downstream tasks. This can be seen with CogVideoX, which provides slight improvements on one of the tasks (*i.e.*, entailment on Human-Hard) while harming the representations on the others (*e.g.*, -1.58 mAP on SSv2-Events).

Are some negative captions challenging? The VideoCon dataset [3] includes negative captions that differ from positive ones by specific types of misalignment, including modifications in actions, attributes, objects, relations, counts, event orders (flipping), and adding hallucinations. Therefore we also analyze whether certain types of captions are particularly challenging for the generators to produce corresponding videos. We achieve this by fine-tuning mPLUG-Owl 7B with synthetic videos specific to each misalignment type. The results averaged over the three video generators are reported in Tab. 2. As shown in the table, different types of misalignment have different impacts on the downstream tasks. For instance, Action is the misalignment that results in the largest overall improvement (*e.g.*, +2.04 mAP on SSv2-Temporal, +0.82% accuracy on ATP-hard), while Flip and Hallucinations misalignments lead to some severe decrease on the VideoCon benchmarks (*e.g.*, -1.12 and -0.61 respectively on VideoCon Human).

We hypothesize that such a performance drop is due to the alignment quality of synthetic videos. To evaluate our hypothesis, we measure the quality of a synthetic video V^s, generated from a caption t^s, as a negative example for the caption t^r as $\bar{f}(V^s, t^s) - \bar{f}(V^s, t^r)$, where $\bar{f}(V, t)$ is computed using an ensemble of VQAScores [29], obtained by averaging the scores from three VQA models [11, 29, 30], *i.e.*, their average likelihood of answering `Yes` to the question: `Does this figure show [t]?` across four uniformly sampled frames from the video. The higher the difference between the two scores, the higher the similarity of the synthetic video to its caption t^s than its negative t^r and, intuitively, the more relevant the synthetic video for the VLA learning process. Fig. 2 shows the distribution of this difference for different types of misalignments. Notably, only Flip and Hallucinations misalignments yield mean differences that are below zero (*i.e.*, -0.03 and -0.05, respectively), while the others are above (*e.g.*, 0.09

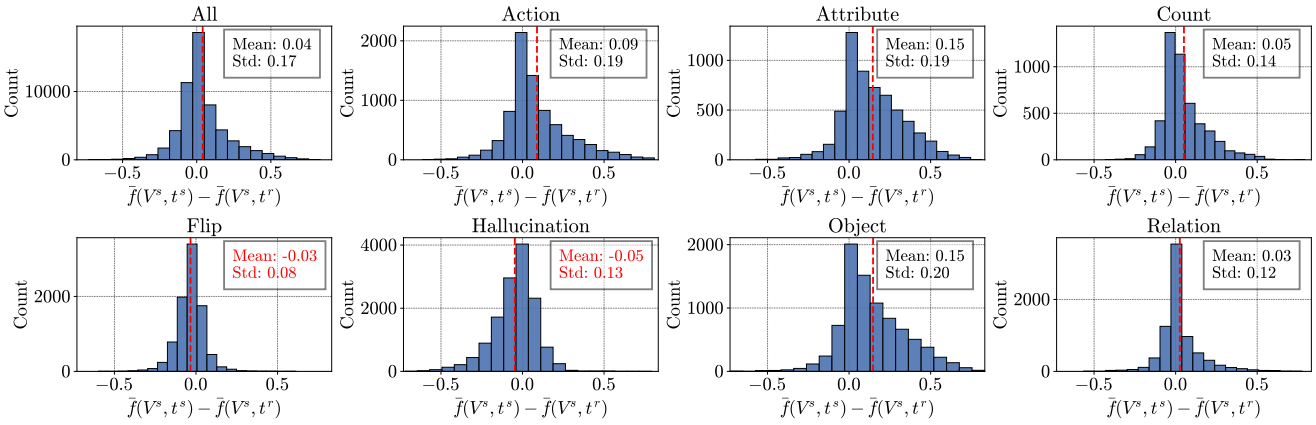

Figure 2. Distribution of the difference between $\bar{f}(V^s, t^s)$ and $\bar{f}(V^s, t^r)$ for each misalignment type, averaged over three text-to-video generators. Misalignment types that result in negative differences (*i.e.*, Flip and Hallucination) are highlighted in red. Best viewed in color.

ACTIONS, 0.15 ATTRIBUTE and OBJECT). This indicates that synthetic videos corresponding to FLIP and HALLUCINATIONS negative captions are not well aligned, which worsens the VLA learning process, as confirmed in Tab. 2.

Finding summary. Our preliminary analysis reveals that: (i) Synthetic videos show potential for enhancing VLA, though improvements are not consistent among different generators. (ii) Different types of misalignment influence various downstream tasks in distinct ways. (iii) Synthetic videos that align closer to the positive captions of real videos rather than the negative captions result in poor training samples, which negatively impact learning.

5. SYNVITA

As shown in the previous section, some generated videos are closer to real captions than their target ones (Fig. 2). This contradicts a key assumption of Eq. (3): that generated videos fully represent the content described by their input caption t^s. This often happens due to semantic inconsistency, *i.e.*, generated videos fail to follow the semantic instruction given by the input text [4, 26]. Such synthetic videos introduce noisy supervision signals, leading to degraded VLA performance (Tab. 2) [32]. Moreover, even semantically consistent synthetic videos may be distinguished using visual differences (*e.g.*, artifacts [38, 52]) rather than intended semantic ones. In this work, we propose a model-agnostic method to better use SYNTHETIC VIDEOS FOR VIDEO-TEXT ALIGNMENT (SYNVITA), modeling them via two strategies: alignment-based weighting and semantic consistency regularization (see Fig. 3).

Alignment-based weighting. To mitigate the impact of harmful synthetic videos and maximize the impact of valuable ones, we weigh the importance of each video based on a scoring criterion ϕ. Given a synthetic video V^s, its corresponding caption t^s and the real counterpart t^r, ϕ maps them to a binary score in $[0, 1]$ depending on their level of alignment, *i.e.*, $\phi : \mathcal{V} \times \mathcal{T} \times \mathcal{T} \rightarrow [0, 1]$. A simple choice for ϕ is to directly use the alignment scores given by our model f. In this case, $\phi(V^s, t^s, t^r) = f(V^s, t^s)$. However, this might ignore the cases where, erroneously, $f(V^s, t^r) > f(V^s, t^s)$, *i.e.*, the generated video is closer to the real caption t^r than to the target one t^s. This phenomenon frequently happens (*i.e.*, Fig. 2), due to *e.g.*, wrong attribute/action binding [23]. For instance, if we ask the model to generate *a horse watching a person running*, it may erroneously generate *a person watching a horse running*, swapping the two actions. As shown in Sec. 4, this type of mistakes harms the learning of f and its capability to distinguish fine-grained details.

Thus, we define ϕ to account for how well the generated video V_i^s represents t_i^s in comparison to its real, negative, counterpart t_i^r, defining the weight for a synthetic video as:

$$\omega_i^\phi = \phi(V_i^s, t_i^s, t_i^r) = \max(0, \bar{f}(V_i^s, t_i^s) - \bar{f}(V_i^s, t_i^r)) \quad (4)$$

where $\bar{f}$ is an ensemble of VQAScores, as in Sec. 4. Note that the more the video is aligned with the target text w.r.t. its negative one, the higher its weight from Eq. (4).

Given this scoring criterion, we define a loss function on synthetic videos, where ϕ acts as a dynamic weight giving higher relevance to videos better aligned with text:

$$\mathcal{L}_{\text{syn}}^\phi = -\sum_{i=1}^{n} \omega_i^\phi \cdot \left(\log f\left(V_i^s, t_i^s\right) + \log\left(1 - f\left(V_i^s, t_i^r\right)\right)\right). \quad (5)$$

Semantic consistency regularization. A positive aspect of having synthetic videos for negative textual inputs is that we can make the model focus on the semantic changes between videos rather than those in appearance. Suppose we are given a text t^r, its negative version t^s, a real video V^r,

Figure 3. Overview of **SYNVITA**. Given a real video V^r with its description t^r and a negative caption t^s (generated by an LLM), we first generate a synthetic video V^s based on t^s. We weigh the importance of each video using the scoring criterion ϕ. We also find the shared semantic between t^r and t^s using the longest common subsequence, obtaining t'. We train f_θ to respond with `Yes` if the input video matches its description and `No` otherwise. Additionally, we encourage the model to focus on the semantic difference between real and synthetic videos, instead of the appearance difference, using their shared semantic (*i.e.*, t').

and its generated negative version V^s. If the difference between t^r and t^s is fine-grained, it will focus on specific properties of the video (*e.g.*, action, temporal order, etc.). This implies that the two texts share most of the content *but* for those fine-grained characteristics. We can thus define a text t', whose semantic is shared between t^r and t^s, thus not being specific to V^r or V^s. We achieve this by finding the intersection between the two texts via longest-common subsequence [20], *i.e.*, $t' = \text{LCS}(t^r, t^s)$[2].

Note that t' has a specific property: given the real (synthetic) video, t' is a less accurate description than the original caption t^r (t^s), but a better one than the negative t^s (t^r). Ideally, our model should capture this relationship, modeling t' as semantically closer to the video than its negative caption, but farther w.r.t. its positive. We can achieve this by computing a triplet loss, defined as:

$$\mathcal{L}^{\phi}_{\texttt{scr}} = \sum_{i=1}^{n} \sum_{z \in \{s,r\}} \omega_i^{\phi} \cdot \big(\max\left(0, \gamma + f(V_i^z, t_i') - f(V_i^z, t_i^z)\right) + \max\left(0, \gamma + f(V_i^z, t_i^{\bar{z}}) - f(V_i^z, t_i')\right)\big). \tag{6}$$

where the margin term γ enforces the desired separation between the alignment probabilities, and when $z = r$, $\bar{z} = s$ and vice versa. The first term promotes better alignment of the positive caption w.r.t. the generic caption t' and the second promotes better alignment of the latter w.r.t. the negative caption. This encourages f to focus on the semantic differences between the two visual inputs, ignoring their differences in appearance due to the synth-to-real gap.

Full objective. Considering all learning objectives together, we obtain the following final function:

$$\mathcal{L} = \mathcal{L}_{\texttt{real}} + \mathcal{L}^{\phi}_{\texttt{syn}} + \lambda_{\texttt{scr}} \cdot \mathcal{L}^{\phi}_{\texttt{scr}}. \tag{7}$$

where $\lambda_{\texttt{scr}}$ is a hyperparameter that regulates the losses. We use Eq. (7) to learn the set θ of parameters in f. Remarkably, our framework has only two hyperparameters, *i.e.*, the margin γ of $\mathcal{L}^{\phi}_{\texttt{scr}}$ and the weight $\lambda_{\texttt{scr}}$ of $\mathcal{L}^{\phi}_{\texttt{scr}}$.

[2]Note that, in practice, t' is not implemented via token removal but via attention-level masking.

6. Experiments

Datasets. For training SYNVITA, we use the VideoCon dataset [3], which includes temporally-challenging video-text triplets from MSR-VTT [57], VATEX [48], and TEMPO [18] for two tasks: *Video-Language Entailment (VLE)* and *Natural Language Explanation (NLE)*. In VLE, the model outputs a score of 1 if the video entails the description and 0 otherwise, while in NLE, it outputs the explanation of the differences between a video and a caption. For each negative caption in the VideoCon VLE training set, we generate a corresponding video using three text-to-video models: CogVideoX [58], LaVie [49], and VideoCrafter2 [7]. The inference configurations for these models and examples of generated videos are in the Supp. Mat.

For evaluation, we use the VideoCon VLE test sets: (i) **VideoCon (LLM)**, with 27K video-text pairs from the same source datasets; (ii) **VideoCon (Human)**, with 570 pairs from ActivityNet [6] and human annotated negative captions; and (iii) **VideoCon (Human-Hard)**, a subset of 290 temporally challenging instances. Following Bansal et al. [3], we also evaluate our model on various downstream tasks: (i) text-to-video retrieval with **SSv2-Temporal** [42], which includes 18 action classes, each with 12 videos (in total 216 videos), requiring temporal understanding; (ii) **SSv2-Events** [2], with 49 action classes, each with 12 videos, featuring multi-event actions; and (iii) video question answering on **ATP-Hard** [5], a subset of questions of NExT-QA [54] that require causal and temporal understanding of videos. We measure the performance using AUC for entailment, mAP for retrieval, and accuracy for VQA. Additional details on the datasets are in the Supp. Mat.

Implementation details. We implement SYNVITA on two

video LLMs, mPLUG-Owl 7B [60] and Video-LLaVA [27], trained on 4 NVIDIA A100 GPUs. Both models share most of the hyperparameters with VideoCon [3] to ensure a fair comparison, and fine-tune the projection layers of the attention blocks of the LLM with low-rank adaptation (LoRA) [22], with $r = 32$, $\alpha = 32$, and dropout $= 0.05$. For both models, we set γ to 0.2, while $\lambda_{\texttt{scr}}$ to 10^{-2} for mPLUG-Owl 7B and 1.0 for Video-LLaVA. Other implementation details can be found in the Supp. Mat.

Baselines. We compare SYNVITA (mPLUG-Owl 7B) and SYNVITA (Video-LLaVA) against two sets of models. The first set includes off-the-shelf VLMs such as VideoCLIP [55], ImageBind (Video-Text) [16], End-to-End VNLI [59], mPLUG-Owl 7B [60], and Video-LLaVA [27], as well as models fine-tuned for improved understanding of actions and event order, *i.e.*, VFC [35] and TACT [2]. The second set consists of models trained on video-text triplets from the VideoCon dataset, namely VideoCon (mPLUG-Owl 7B) and VideoCon (Video-LLaVA) [3]. Additional details on the baselines are in the Supp. Mat.

6.1. Comparison with state of the art

Tab. 3 presents the results of our comparison on the VideoCon evaluation sets and the downstream tasks. Overall, our proposed method outperforms all previous baselines in five tasks out of six. For the entailment task, on the VideoCon Human dataset SYNVITA (Video-LLaVA) improves its counterpart VideoCon (Video-LLaVA), trained without synthetic video-caption pairs, by 0.77%, and achieves a 1.12% improvement on its temporally challenging subset, Human-Hard. Similarly, SYNVITA (mPLUG-Owl 7B) shows a 0.32% improvement on the VideoCon Human dataset. As expected from Sec. 3, on the VideoCon (LLM) test set, both SYNVITA (Video-LLaVA) and SYNVITA (mPLUG-Owl 7B) underperform compared to their counterparts without synthetic videos, due to the similar distribution of negatives w.r.t. those present in the training set. Thus, synthetic pairs harm the performance in this setting.

For text-to-video retrieval tasks, SYNVITA (mPLUG-Owl 7B) outperforms VideoCon (mPLUG-Owl 7B) by 4.32% on SSv2-Temporal and 2.17% on SSv2-Events. Similarly, SYNVITA (Video-LLaVA) shows improvements of 0.33% on SSv2-Temporal and 1.20% on SSv2-Events compared to VideoCon (Video-LLaVA). These results suggest that our model is model-agnostic and more effective at ranking similar but semantically different text descriptions than the baseline, which does not associate corresponding video data with negative captions. Finally, for the challenging video question-answering task on the ATP-Hard dataset, models fine-tuned with only textual negatives see performance drops or minimal improvement compared to their non-finetuned version. Despite this, SYNVITA (mPLUG-Owl 7B) improves upon VideoCon (mPLUG-Owl 7B) by 1.85%, and SYNVITA (Video-LLaVA) shows a 1.12% improvement over VideoCon (Video-LLaVA). We show qualitative results of SYNVITA in the Supp. Mat.

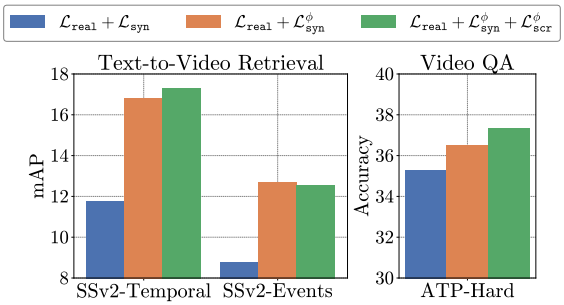

Figure 4. Ablation study on the proposed losses.

6.2. Ablation study

In this section, we analyze the components of SYNVITA considering the mPLUG-Owl 7B version. We first examine the different parts of our learning objective. We then show the benefits of alignment-based weighting over fixed weights and of different alignment-based scoring criteria. Finally, we show the effect of using different text-to-video models for generating synthetic videos. Additional results on other designs are present in the Supp. Mat.

Learning objectives. We first analyze the effectiveness of the two proposed components in our learning objective: the alignment-based loss function $\mathcal{L}^{\phi}_{\texttt{syn}}$ (Eq. (5)) and the semantic consistency regularization $\mathcal{L}^{\phi}_{\texttt{scr}}$ (Eq. (6)). As shown in Fig. 4, excluding $\mathcal{L}^{\phi}_{\texttt{syn}}$ leads to a drop in performance (blue vs. orange bar). Without this loss, the objective is solely the traditional language modeling loss. As a result, synthetic videos that are not aligned with their captions introduce a noisy training signal. Adding $\mathcal{L}^{\phi}_{\texttt{scr}}$ (green bar), further boosts the performance on 2/3 datasets, suggesting that the model better captures the video semantics. As our model is trained on triplets with single-event differences [3], $\mathcal{L}^{\phi}_{\texttt{scr}}$ is less effective for SSv2-Events, where captions involve multiple events. However, current open-source video generators struggle to generate multi-event videos.

Alignment-based weighting strategy. In this section, we evaluate our alignment-based weighting strategy (*i.e.*, Eq. (4)) against other alternatives, reporting the results in Tab. 4. As a reference, row (1) reports the results of a fixed weight (*i.e.*, 1) for all synthetic videos. Assigning weights based only on alignment with the target text (*i.e.*, $f(V^s, t^s)$) improves performance on retrieval (*e.g.*, +0.92 mAP on SSv2-Events) but degrades performance on others (*e.g.*, on ATP-Hard, -0.28%), as it overlooks cases where synthetic videos align more with real captions. In row (3), we multiply the synthetic scores by the inverse similarity with the real counterpart (*i.e.*, $(1 - f(V^s, t^r))$). Introducing the real captions into the score improves the results in var-

	Video-language Entailment (VideoCon)			Text-to-Video Retrieval		Video QA
	LLM	Human	Human-Hard	SSv2-Temporal	SSv2-Events	ATP-Hard
VideoCLIP [55]	53.2	47.3	47.5	9.8	6.4	23.4
ImageBind (Video-Text) [16]	57.1	65.2	63.0	10.5	5.5	25.4
TACT [2]	-	-	-	-	7.8	27.6
VFC [35]	-	-	-	-	-	31.4
End-to-End VNLI [59]	67.0	72.4	65.0	14.6	10.4	39.0
mPLUG-Owl 7B [60]	57.24	67.02	64.39	11.08	6.75	37.96
Video-LLaVA [27]	62.98	70.37	65.99	11.64	7.11	38.56
VideoCon (mPLUG-Owl 7B)* [3]	**88.39**	77.16	74.76	13.00	10.37	35.46
VideoCon (Video-LLaVA)	85.86	80.09	75.74	19.77	10.01	38.76
SynViTA (mPLUG-Owl 7B)	86.45	77.48	74.54	17.32	**12.54**	37.31
SynViTA (Video-LLaVA)	85.43	**80.86**	**76.86**	**20.10**	11.21	**39.88**

Table 3. Comparison of SynViTA with both discriminative and generative VLMs. For the video-language entailment task, we report AUC-ROC, for zero-shot text-to-video retrieval, we report mAP, and for video question-answering, we report accuracy. * indicates our reproduced results using the mPLUG-Owl 7B model checkpoint released in the original VideoCon repository.

Alignment-based weighting strategy	Video-Language Entailment (VideoCon)			Text-to-Video Retrieval		Video QA
	LLM	Human	Human-Hard	SSv2-Temporal	SSv2-Events	ATP-Hard
1) Fixed weighting-1.00	83.95	76.91	**75.05**	12.54	8.48	36.23
2) $\bar{f}(V^s, t^s)$	84.87	76.46	74.12	13.43	9.40	35.95
3) $\bar{f}(V^s, t^s) \cdot (1 - \bar{f}(V^s, t^r))$	85.57	76.79	74.11	15.52	10.74	36.06
4) $\mathbb{1}\left[\bar{f}(V^s, t^s) > \bar{f}(V^s, t^r)\right]$	84.88	76.79	73.17	14.17	10.38	37.15
5) $\max(0, \bar{f}(V^s, t^s) - \bar{f}(V^s, t^r))$	**86.45**	**77.48**	74.54	**17.32**	**12.54**	**37.31**

Table 4. Results of the ablation study on the weighting strategy for the synthetic videos in the objective function.

	Text-to-Video Generator				
	None	CogVideoX	LaVie	VideoCrafter2	All
LLM	**88.39**	86.45	86.45	86.43	85.82
Human	77.16	77.48	**77.51**	77.48	77.15
Human-Hard	**74.76**	74.54	74.73	74.74	73.79
SSv2-Temporal	13.00	**17.32**	15.98	15.47	14.06
SSv2-Events	10.37	**12.54**	12.36	11.72	10.90
ATP-Hard	35.46	**37.31**	36.55	36.50	36.44

Table 5. Ablation study on varying the text-to-video model.

ious settings, especially on retrieval, achieving +2.98 mAP on SSv2-Temporal, and +2.26 mAP on SSv2-Events. As an alternative, row (4) considers weighing all synthetic videos as 1 if they are closer to their target caption than the real one. This strategy shows a general degradation w.r.t. the previous, except for ATP-Hard (+1.9%). This denotes that a soft-weighting scheme is still more effective as it accounts for different levels of semantic fidelity across videos. Our proposed strategy (row (5)) combines the advantages of the two, enforcing that synthetic videos are truly negative examples, *i.e.*, being more similar to their caption than the original one of the real videos. This strategy obtains the highest results in almost all settings. For Video-LLaVA, we use (3) as it performs slightly better.

Text-to-video generators. We analyze this aspect by comparing three text-to-video generators when used with our method: CogVideoX [58], LaVie [49], and VideoCrafter2 [7]. As shown in Tab. 5, SynViTA (mPLUG-Owl 7B) fine-tuned on videos generated by CogVideoX outperforms the other alternatives across all downstream tasks and achieves comparable results on the video-language entailment task. While it can be challenging to determine *a-priori* the optimal generator for a downstream task, one possibility could be to generate videos from multiple generators and let the model filter them. Using all generated videos performs better than using none on the downstream tasks (*e.g.*, +1.06% on SSv2-Temporal), but underperforms CogVideoX (*e.g.*, -3.26% on SSv2-Temporal). This is likely due to the high synth-to-real video ratio, introducing a significant domain shift that requires careful handling. Nevertheless, we expect that the better the text-to-video models released, the more beneficial they will be for SynViTA.

7. Conclusion

In this work, we explored whether videos generated by text-to-video models can help learning a better video-language alignment (VLA) model. Our initial analysis shows that synthetic videos can boost performance on certain downstream tasks, but harm others. We attribute this to (i) semantic inconsistency, as synthetic videos may not follow the input text, and (ii) appearance bias, where the model focuses on visual differences in the videos rather than semantic differences. To address these limitations, we introduced SynViTA, the first VLA method exploiting synthetic videos. SynViTA includes an alignment-based sample weighting strategy to mitigate noisy video generations and a semantic consistency regularization to make the model focus on semantic, rather than visual, differences. SynViTA outperforms baselines that do not use synthetic videos across different video LLMs on five out of six tasks, demonstrating its potential to improve VLA across diverse models.

Acknowledgments. This work was sponsored by EU ISFP PRECRISIS (ISFP2022-TFI-AG-PROTECT-02-101100539), PNRR ICSC National Research Centre for HPC, Big Data and Quantum Computing (CN00000013), Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007), the FAIR - Future AI Research (PE00000013), funded by NextGeneration EU, and EU PATTERN (Project No. 101159751). We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

References

[1] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv*, 2023. 1

[2] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *CVPR*, 2023. 2, 3, 6, 7, 8, 1

[3] Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. In *CVPR*, 2024. 1, 2, 3, 4, 6, 7, 8

[4] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv*, 2024. 5

[5] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *CVPR*, 2022. 2, 4, 6, 1

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 6, 1

[7] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 2, 3, 4, 6, 8, 1

[8] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org*, 2023. 2

[9] Sanjoy Chowdhury, Sayan Nag, and Dinesh Manocha. Apollo: Unified adapter and prompt learning for vision language models. In *EMNLP*, 2023. 3

[10] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv*, 2024. 1, 2

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven CH Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv*, 2023. 4, 2, 3

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 1, 2

[13] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *NeurIPS*, 2024. 3

[14] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. In *arXiv*, 2021. 2

[15] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *TPAMI*, 2024. 1

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 7, 8, 2

[17] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv*, 2022. 2, 3

[18] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *EMNLP*, 2018. 6, 1

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2

[20] Daniel S Hirschberg. Algorithms for the longest common subsequence problem. *JACM*, 1977. 6

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3

[22] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 7, 1

[23] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*, 2023. 5

[24] Diederik P Kingma. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1, 2

[25] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *ACL*, 2023. 1

[26] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *CVPRW*, 2024. 2, 5, 3

[27] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv*, 2023. 2, 7, 8, 1

[28] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv*, 2023. 1, 2

[29] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv*, 2024. 1, 2, 3, 4

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 4, 2, 3

[31] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv*, 2024. 1, 2

[32] Haoyu Lu, Mingyu Ding, Nanyi Fei, Yuqi Huo, and Zhiwu Lu. Lgdn: Language-guided denoising network for video-language modeling. *NeurIPS*, 2022. 5

[33] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 2022. 2

[34] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *CVPR*, 2024. 2, 3

[35] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *ICCV*, 2023. 1, 2, 7, 8

[36] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *NeurIPS*, 2024. 2, 3

[37] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. *arXiv*, 2023. 2

[38] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. 5

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[40] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *AAAI*, 2024. 3

[41] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In *CVPR*, 2023. 2

[42] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *WACV*, 2021. 2, 3, 6, 1

[43] Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *CVPR*, 2022. 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv*, 2020. 3

[45] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *NeurIPS*, 2024. 2, 3

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv*, 2023. 2

[47] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *AAAI*, 2024. 1

[48] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 6, 1

[49] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv*, 2023. 2, 3, 4, 6, 8, 1

[50] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *NeurIPS*, 2022. 2

[51] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *NeurIPS*, 2024. 2

[52] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv*, 2024. 2, 5

[53] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu

Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv*, 2024. 2

[54] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 6, 1

[55] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv*, 2021. 1, 2, 7, 8

[56] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *ICML*, 2023. 2

[57] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 6, 1

[58] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv*, 2024. 2, 3, 4, 6, 8, 1

[59] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *NeurIPS*, 2024. 7, 8, 1, 2

[60] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv*, 2023. 2, 3, 7, 8, 1

[61] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022. 2

[62] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into clip latent space for video anomaly recognition. *arXiv*, 2023. 2

[63] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *CVPR*, 2024. 2

[64] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023. 2

[65] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv*, 2023. 2, 3

[66] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv*, 2023. 2