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

Supplementary Material

In this supplementary material, we provide additional details about our implementation (Sec. A), the training and evaluation datasets (Sec. B), performance metrics (Sec. C), and baseline models (Sec. D). We also provide further analyses of design choices in Sec. E, along with the inference configuration of text-to-video models and examples of generated videos in Sec. F. Finally, we describe the limitations of our method in Sec. G and present qualitative results on evaluation datasets in Sec. H.

A. Implementation details

We implement SYNVITA on two video large language models (LLMs): mPLUG-Owl 7B [60] and Video-LLaVA [27], trained on 4 NVIDIA A100 GPUs with 64GB memory each. We train the same layers as in VideoCon [3] to ensure a fair comparison. Specifically, we fine-tune the projection layers of the attention blocks of the LLM using low-rank adaptation (LoRA) [22] with parameters $r = 32$, $\alpha = 32$, and dropout $= 0.05$. We train the model for 1 epoch for both mPLUG-Owl 7B and Video-LLaVA. Due to memory constraints, we adjust the batch sizes for each model: mPLUG-Owl 7B uses a batch size of 8, and Video-LLaVA uses a batch size of 4. Both models are trained using the Adam optimizer [24], with a linear warmup of 200 steps, a cosine annealing schedule, and learning rates of 10^{-4} for mPLUG-Owl 7B and 5×10^{-5} for Video-LLaVA. We empirically set the margin term γ to 0.2 for both models and the consistency weight $\lambda_{\texttt{scr}}$ to 10^{-2} for mPLUG-Owl 7B and 1.0 for Video-LLaVA.

B. Details about 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. In NLE, the model outputs an explanation of the difference between a video and a caption.

For training, we use the same training split of VideoCon [3], containing about 108K video-text triplets for both tasks. The VideoCon dataset includes negative captions generated to represent seven types of semantically plausible misalignments: *object, action, attribute, counting, relation, hallucination*, and *event order flip*. For each negative caption, we generate a corresponding video using three open-source text-to-video generation models: CogVideoX [58], LaVie [49], and VideoCrafter2 [7], resulting in a total of 173,337 generated videos.

For evaluation, we use the VLE test sets of VideoCon, which include i) **VideoCon (LLM):** 27K video-text pairs from the same datasets used for training, with negative captions generated by an LLM; ii) **VideoCon (Human):** 570 pairs from ActivityNet [6], where negative captions are manually annotated; iii) **VideoCon (Human-Hard):** a subset of VideoCon (Human) with 290 pairs labeled as temporally challenging (*i.e.*, each video frame does not entail the caption) by an image-text alignment model [59].

Following [3], we also evaluate our model on downstream tasks that require temporal understanding. Specifically, we test on: i) text-to-video retrieval using **SSv2-Temporal** [42] and **SSv2-Events** [2]; and ii) on video question answering using **ATP-Hard** [5]. The SSv2-Temporal dataset includes 18 action classes, each with 12 matching videos (in total 216 videos), featuring actions such as *Moving [something] and [something] away from each other*. The SSv2-Events dataset contains 49 action classes with 12 videos per class. SSv2-Events focuses on templates involving multiple verbs, which indicates various events within a video, such as *Pouring [something] into [something] until it overflows*. Finally, ATP-Hard [5] is a subset of questions of the NExT-QA benchmark [54] that require causal and temporal understanding of videos where, for each question, there are 5 possible answers.

C. Details about performance metrics

We follow the evaluation protocol established by Bansal et al. [3] and measure the video-language entailment performance using the area under the receiver operating characteristic curve (AUC ROC) on the VLE test sets of VideoCon. For text-to-video retrieval, we compute video-language alignment scores between each action class and the candidate videos, rank the videos, and report the mean Average Precision (mAP). For video question answering, we use statements generated by Bansal et al. [3] via the PaLM-2 API [1]. We compute the alignment scores of these statements with the input video, select the highest-ranking statement, and report the accuracy.

D. Details about baselines

We implement both VIDEOCON baselines and our SYNVITA with two video large language models: mPLUG-Owl 7B [60] and Video-LLaVA [27]. *mPLUG-Owl 7B* employs a CLIP ViT-L/14 [12] visual encoder to extract features from 32 uniformly sampled video frames. These features are processed by a visual abstractor module, which in-

	Video-Language Entailment (VideoCon)		
	LLM	Human	Human-Hard
InstructBLIP [29]	64.97	73.37	66.87
LLaVA-1.5 [29]	64.54	69.84	63.67
CLIP-FlanT5 [29]	65.59	74.41	67.92
VQAScores [29]	66.60	74.60	67.85
SynViTA (mPLUG-Owl 7B)	**86.45**	77.48	74.54
SynViTA (Video-LLaVA)	85.43	**80.86**	**76.86**

Table 6. Comparison of SynViTA with image-text alignment models [11, 29, 30].

cludes additional temporal query tokens for temporal modeling, to compress the visual information into a fixed number of learnable tokens. The resulting tokens are combined with tokenized textual queries and provided as input into LLaMA-7B [46], serving as the large language model. *Video-LLaVA* employs LanguageBind encoders [66], initialized from CLIP ViT-L/14 [12], to map 8 uniformly sampled video frames into the textual feature space of LanguageBind [66]. A 2-layer fully connected network processes these features, which are then combined with tokenized textual queries and fed as input to Vicuna-7B v1.5 [8], serving as the large language model.

For the baseline VideoCon (mPLUG-Owl 7B), we report results in the main manuscript using the checkpoint from the latest version of the official VideoCon repository, which includes the corrected LoRA α parameter introduced in commit `9a69520`.

For the baseline VideoCon (Video-LLaVA), we report results obtained by fine-tuning the Video-LLaVA model on the VideoCon dataset using the same trained layers and hyperparameters as in VideoCon [3]. Specifically, we train the model for 2 epochs, fine-tuning the projection layers of the LLM's attention blocks using LoRA with parameters $r = 32$, $\alpha = 32$, and dropout $= 0.05$. We use a batch size of 16, the Adam [24] optimizer, a linear warmup of 200 steps, a cosine annealing learning rate schedule, and a learning rate of 10^{-4}.

Finally, apart from the off-the-shelf versions of mPLUG-Owl 7B and Video-LLaVA, for which we performed evaluation on the evaluation sets, the results of all other baselines (*i.e.*, VideoCLIP [55], ImageBind (Video-Text) [16], End-to-End VNLI [59], VFC [35], and TACT [2]) are taken from the VideoCon paper [3].

E. Additional analysis

In this section, we analyze the models used to estimate alignment scores for our alignment-based weighting strategy, the hyperparameters of our model, *i.e.*, the margin and the weight of the semantic consistency regularization, and SynViTA's application to specific types of misalignment.

Ensemble for image-text alignment models. Our alignment-based weighting strategy relies on alignment scores between synthetic videos and captions. We thus analyze the performance of the state-of-the-art approach [29] for computing these scores on the video-language entailment task, which is the closest task to image-text alignment. Specifically, we evaluate three multimodal LLMs: LLaVA-1.5 [30], InstructBLIP [11], and CLIP-FlanT5 [29]. In addition to using them individually, we include an ensemble version referred to as VQAScores [29], which averages the scores of the image-text alignment models.

As shown in Table 6, the ensemble VQAScores generally outperforms individual models, as videos that consistently receive high alignment scores across all models are more likely to be consistent with their textual description. In SynViTA, we use this ensemble to evaluate the quality of the alignment, as this offers more precise scores than individual models. For reference, we also report the results obtained by SynViTA (mPLUG-Owl 7B) and SynViTA (Video-LLaVA). These results highlight a performance gap between models fine-tuned on the video-language entailment task and off-the-shelf multimodal LLMs. Specifically, SynViTA (Video-LLaVA), which relies on a smaller LLM (7B, compared to the 13B for LLaVA-1.5, and 11B for InstructBLIP and CLIP-FlanT5), improves the ensemble's performance by 18.83%, 6.26%, and 9.01% on the

Margin	Video-Language Entailment (VideoCon)		
	LLM	Human	Human-Hard
0.2	**86.14**	77.25	**73.95**
0.4	86.04	**77.39**	73.71
0.6	86.03	77.17	73.72
0.8	85.99	76.78	73.39

Table 7. Results of varying the margin term used for semantic consistency regularization.

Weight	Video-Language Entailment (VideoCon)		
	LLM	Human	Human-Hard
10^{-3}	**86.45**	**77.48**	74.53
10^{-2}	**86.45**	**77.48**	**74.54**
10^{-1}	86.40	77.34	74.29
1.0	86.14	77.25	73.95

Table 8. Results of varying the weight term used for semantic consistency regularization.

	Misalignment Type			
	Action	Action (SynViTA)	Hall.	Hall. (SynViTA)
LLM	86.10	85.80 (↓ 0.30)	85.46	**86.45** (↑ 0.99)
Human	77.43	**77.86** (↑ 0.43)	76.55	77.03 (↑ 0.48)
Human-Hard	74.83	**74.96** (↑ 0.13)	74.77	74.74 (↓ 0.03)
SSv2-Temporal	15.04	**15.41** (↑ 0.37)	13.89	14.89 (↑ 1.00)
SSv2-Events	10.66	**11.66** (↑ 1.00)	10.14	11.00 (↑ 0.86)
ATP-Hard	36.28	**37.62** (↑ 1.34)	36.37	36.42 (↑ 0.05)

Table 9. Average results of applying SynViTA to specific misalignment types across different text-to-video models. Increases (↑) and decreases (↓) are measured relative to the model "blindly" fine-tuned with synthetic videos of the same misalignment type.

Model	Resolution (W×H)	Length (frames)	FPS (frames/s)	Guidance Scale	Sampling Steps	Noise Scheduler	Generation Time (s)
CogVideoX [58]	720×480	49	8	6.0	50	DDIM [44]	∼75
LaVie [49]	512×320	32	8	7.5	50	DDPM [21]	∼20
VideoCrafter2 [7]	512×320	32	8	12.0	50	DDIM [44]	∼109

Table 10. Inference configurations for text-to-video generators.

VideoCon evaluation datasets. This supports the finding that off-the-shelf multimodal LLMs often lack robustness to fine-grained caption manipulations [26]. Nevertheless, we can use their prior knowledge to estimate the semantic consistency of the generated videos.

Margin term γ of $\mathcal{L}^{\phi}_{\texttt{scr}}$. The margin term γ of the semantic consistency regularization loss controls the desired separation between alignment probabilities. We analyze its effect by fixing the loss weight to 1 and present the results in Tab. 7. As can be seen, setting γ to 0.2 for mPLUG-Owl 7B achieves the best performance on two out of three video-language entailment datasets.

Weight $\lambda_{\texttt{scr}}$ of $\mathcal{L}^{\phi}_{\texttt{scr}}$. The weight $\lambda_{\texttt{scr}}$ of the semantic consistency regularization regulates its contribution to the overall objective. We analyze the effect of varying this hyperparameter and report the results in Tab. 8. Setting the value to 10^{-2} for mPLUG-Owl 7B achieves the highest performance across all video-language entailment datasets.

SynViTA on specific types of misalignment. Tab. 2 shows how different types of misalignment affect downstream tasks. A natural question is how SynViTA performs when applied to a specific misalignment. To explore this, we select one misalignment with a positive mean (*i.e.*, ACTION) and one with a negative mean (*i.e.*, HALLUCINATION) from Fig. 2. We then apply SynViTA to videos of these categories, generated by three video generators, and report the average results in Tab. 9. The results show that SynViTA improves performance on 5 out of 6 tasks (*e.g.*, +1% and +0.86% on SSv2-Events) compared to the model "blindly" fine-tuned on the same synthetic videos.

F. Analysis of text-to-video generators

For each negative caption in the VideoCon dataset [3], we generate a corresponding video using three open-source text-to-video generation models: CogVideoX [58], LaVie [49], and VideoCrafter2 [7], resulting in a total of 173,337 generated videos. The inference configurations used for these models are reported in Tab. 10. For each model, we use the default configuration available at the time of cloning the respective GitHub repository, with the following exceptions: for LaVie and VideoCrafter2, we set the number of frames to 32 and the frames per second to 8 to ensure that the generated videos have a duration of 4 seconds (the longest we can obtain with the available models). As shown in the table, for CogVideoX, we generate longer videos (approximately 6.125 seconds) with a higher resolution (720×480 pixels) compared to the 4-second videos and 512×320 pixel resolution of LaVie and VideoCrafter2. Using NVIDIA A100 GPUs, the average generation times per video are: (1) CogVideoX: 1.51s per step, ∼75s total. (2) LaVie: 0.4s per step, ∼20s total. (3) VideoCrafter2: 2.18s per step, ∼109s total.

Fig. 5 presents examples of videos generated from LLM-generated negative captions, along with alignment scores assigned by image-text alignment methods (*i.e.*, InstructBLIP [11], LLaVA-1.5 [30], and CLIP-FlanT5 [29]). Specifically, we show the alignment of each synthetic video V^s with its corresponding caption t^s, denoted as $f(V^s, t^s)$, and with its real counterpart t^r, denoted as $f(V^s, t^r)$. For the caption *A man talks about a plate of tacos while wearing a sombrero*, all three models generate semantically consistent videos, resulting in higher alignment scores for the input caption than the real counterpart. In contrast, for the caption *A man is holding two large dumbbells which he raises up and down in both hands*, CogVideoX fails to depict the size of the dumbbells. In this case, the alignment between the synthetic video and the corresponding caption is lower than that of the real counterpart for two out of three models. Finally, for the caption *A man is trimming the bottom of a palm tree and then climbs it*, none of the generated videos achieve higher alignment scores with the input caption than with the real caption. This is likely due to the nonsensical nature of the caption generated by the LLM. As this caption belongs to the *event order flip* type of misalignment, if many captions of this type are similarly affected by this issue, it could explain the negative mean alignment difference observed in Fig. 2 for this category of synthetic videos. Such negative mean may also result from artifacts introduced by text-to-video models. This is particularly evident in the video generated by VideoCrafter2, where the model only includes the top of a palm tree to adhere to the input prompt.

G. Limitations

Our method depends on the capability of text-to-video generators, which are constrained to produce short-duration videos, often shorter than real ones. This may contribute to a larger syn-real shift and lead to less temporally challenging synthetic videos, limiting the learning strength of our method. While our work pioneers this research direction, further benefits will come from future advances in high-quality video generation and alignment evaluation.

t^r: a man talks about his plate of tacos

t^s: a man talks about his plate of tacos while wearing a sombrero

	CogVideoX		LaVie		VideoCrafter2	
	$f(V^s, t^s)$	$f(V^s, t^r)$	$f(V^s, t^s)$	$f(V^s, t^r)$	$f(V^s, t^s)$	$f(V^s, t^r)$
InstructBLIP	0.85	0.70	0.94	0.88	0.96	0.89
LLaVA-1.5	0.65	0.60	0.59	0.57	0.60	0.64
CLIP-FlanT5	0.74	0.62	0.86	0.74	0.91	0.83

t^r: Man is holding two small dumbells which he raises up and down in both of his hands.

t^s: Man is holding two large dumbells which he raises up and down in both of his hands.

	CogVideoX		LaVie		VideoCrafter2	
	$f(V^s, t^s)$	$f(V^s, t^r)$	$f(V^s, t^s)$	$f(V^s, t^r)$	$f(V^s, t^s)$	$f(V^s, t^r)$
InstructBLIP	0.77	0.79	0.93	0.93	0.93	0.90
LLaVA-1.5	0.60	0.61	0.58	0.43	0.56	0.30
CLIP-FlanT5	0.85	0.82	0.91	0.49	0.93	0.41

t^r: A man is climbing a palm tree and trimming the bottom of it.

t^s: A man is trimming the bottom of a palm tree and then climbs it

	CogVideoX		LaVie		VideoCrafter2	
	$f(V^s, t^s)$	$f(V^s, t^r)$	$f(V^s, t^s)$	$f(V^s, t^r)$	$f(V^s, t^s)$	$f(V^s, t^r)$
InstructBLIP	0.78	0.82	0.65	0.66	0.42	0.66
LLaVA-1.5	0.53	0.75	0.49	0.72	0.40	0.44
CLIP-FlanT5	0.66	0.90	0.72	0.94	0.59	0.84

Figure 5. Examples of videos generated by three text-to-video models (*i.e.*, CogVideoX, LaVie, and VideoCrafter2) from LLM-generated negative captions, along with alignment scores assigned by different image-text alignment methods (*i.e.*, InstructBLIP, LLaVA-1.5, and CLIP-FlanT5). For each synthetic video V^s and alignment model, we show its alignment with the corresponding caption t^s, denoted as $f(V^s, t^s)$, and with the real caption t^r, denoted as $f(V^s, t^r)$.

H. Qualitative results of SYNVITA

Figs. 6 and 7 show examples of video-language alignment scores assigned by SYNVITA (mPLUG-Owl 7B) and SYNVITA (Video-LLaVA), compared to baselines trained without synthetic videos, for the video-language entailment task on VideoCon LLM and VideoCon Human and Human Hard, respectively. Similarly, Fig. 8 presents alignment scores for the video question answering task on ATP-Hard. Finally, Figs. 9 to 12 show rankings based on video-language alignment scores for the text-to-video retrieval task on SSv2-Temporal and SSv2-Events, using SYNVITA (mPLUG-Owl 7B) and SYNVITA (Video-LLaVA) against the same baselines. Figs. 6 and 7 show some success cases on the video-language entailment task, where SYNVITA assigns higher alignment scores to captions matching the videos compared to their negative counterparts (*e.g.*, it better distinguishes the action of standing from sitting on both videos from VideoCon Human). Fig. 6 (rows 2 and 4) also shows two failure cases where SYNVITA incorrectly associates videos with negative captions, unlike the baseline, which makes the correct associations. In the first case, SYNVITA misjudges the number of children because the second child appears briefly at the end, looks similar to the first, and is never seen together. In the second case, SYNVITA fails to detect that oil, not water, is added to the pan, likely because the oil is already inside the pan when the video starts. For the video question answering task (Fig. 8), it better associates the scenario of a child sitting on its father's stomach versus its shoulders. Finally, on the text-to-video retrieval task, it better recognizes certain actions such as moving relative to the camera (Figs. 9 and 10), rolling something (Fig. 11), and pouring something (Fig. 12).

Figure 6. Examples of video-language alignment scores assigned by SynViTA (mPLUG-Owl 7B) and SynViTA (Video-LLaVA), compared to baselines trained without synthetic videos, for the video-language entailment task on VideoCon LLM. Captions marked with green borders correctly match the input video, while those marked with red borders do not. The models' highest predicted scores are highlighted in violet. If the top prediction corresponds to the caption that correctly describes the video, the row is marked with a checkmark; otherwise, it is marked with a cross.

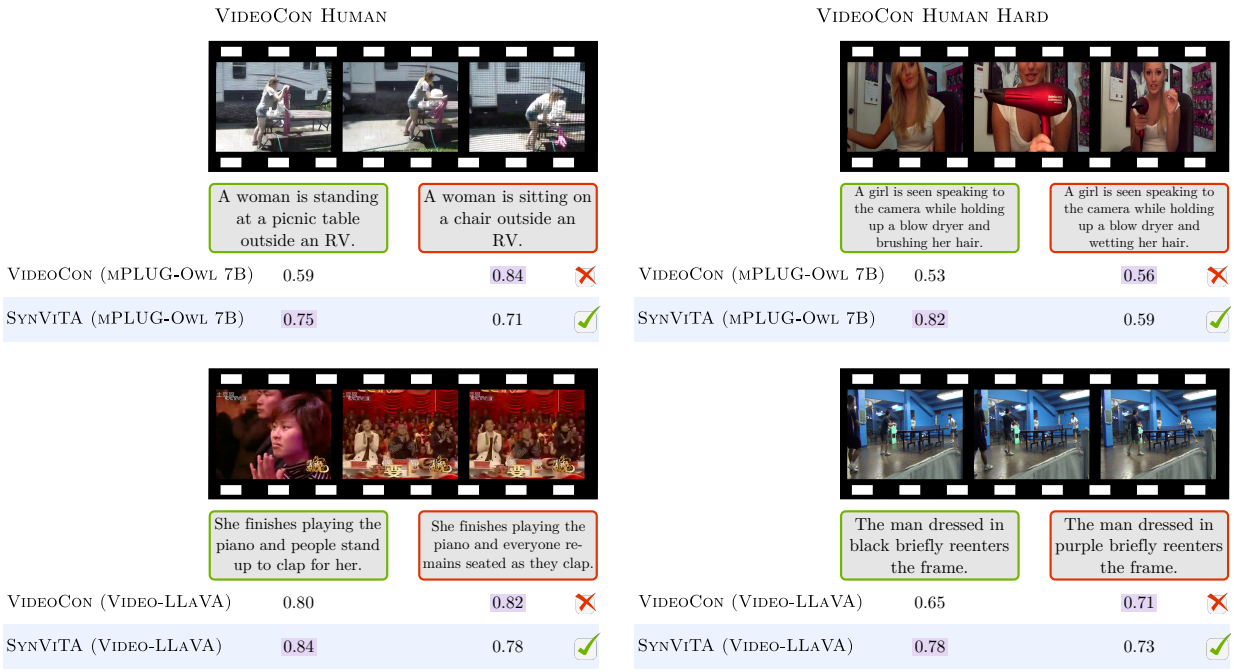

Figure 7. Examples of video-language alignment scores assigned by SynViTA (mPLUG-Owl 7B) and SynViTA (Video-LLaVA), compared to baselines trained without synthetic videos, for the video-language entailment task on VideoCon Human and Human Hard. Captions marked with green borders correctly match the input video, while those marked with red borders do not. The models' highest predicted scores are highlighted in violet. If the top prediction corresponds to the caption that correctly describes the video, the row is marked with a checkmark; otherwise, it is marked with a cross.

ATP-HARD

what did the lady in blue do as the black dog is swimming?

	lady in blue follows behind as the black dog is swimming	lady in blue talks to girl in grey as the black dog is swimming	lady in blue assists it as the black dog is swimming	lady in blue catapults as the black dog is swimming	lady in blue floats away as the black dog is swimming	
VIDEOCON (MPLUG-OWL 7B)	0.73	0.75	0.71	0.65	0.56	✗
SYNVITA (MPLUG-OWL 7B)	0.75	0.73	0.78	0.73	0.73	✓

how is the boy being held while the man is on the swing?

	boy is being held on his back while the man is on the swing.	boy is being held on the man's stomach while the man is on the swing.	boy is being held on his shoulders while the man is on the swing.	boy is being held on his knees while the man is on the swing.	boy is being held on the man's lap while the man is on the swing.	
VIDEOCON (VIDEO-LLAVA)	0.71	0.65	0.73	0.59	0.62	✗
SYNVITA (VIDEO-LLAVA)	0.68	0.71	0.68	0.59	0.68	✓

Figure 8. Examples of video-language alignment scores assigned by SYNVITA (mPLUG-Owl 7B) and SYNVITA (Video-LLaVA), compared to baselines trained without synthetic videos, on the video question answering task for ATP-Hard. Captions marked with green borders are the correct answers. The models' highest predicted scores are highlighted in violet. If the top prediction corresponds to the correct answer, the row is marked with a checkmark; otherwise, it is marked with a cross.

Figure 9. Comparison of rankings based on video-language alignment scores for the text-to-video retrieval task on SSv2-Temporal, using SynViTA (mPLUG-Owl 7B) against the baseline VideoCon (mPLUG-Owl 7B) trained without synthetic videos. Videos marked with green borders correctly match the input text query, while those marked with red borders do not.

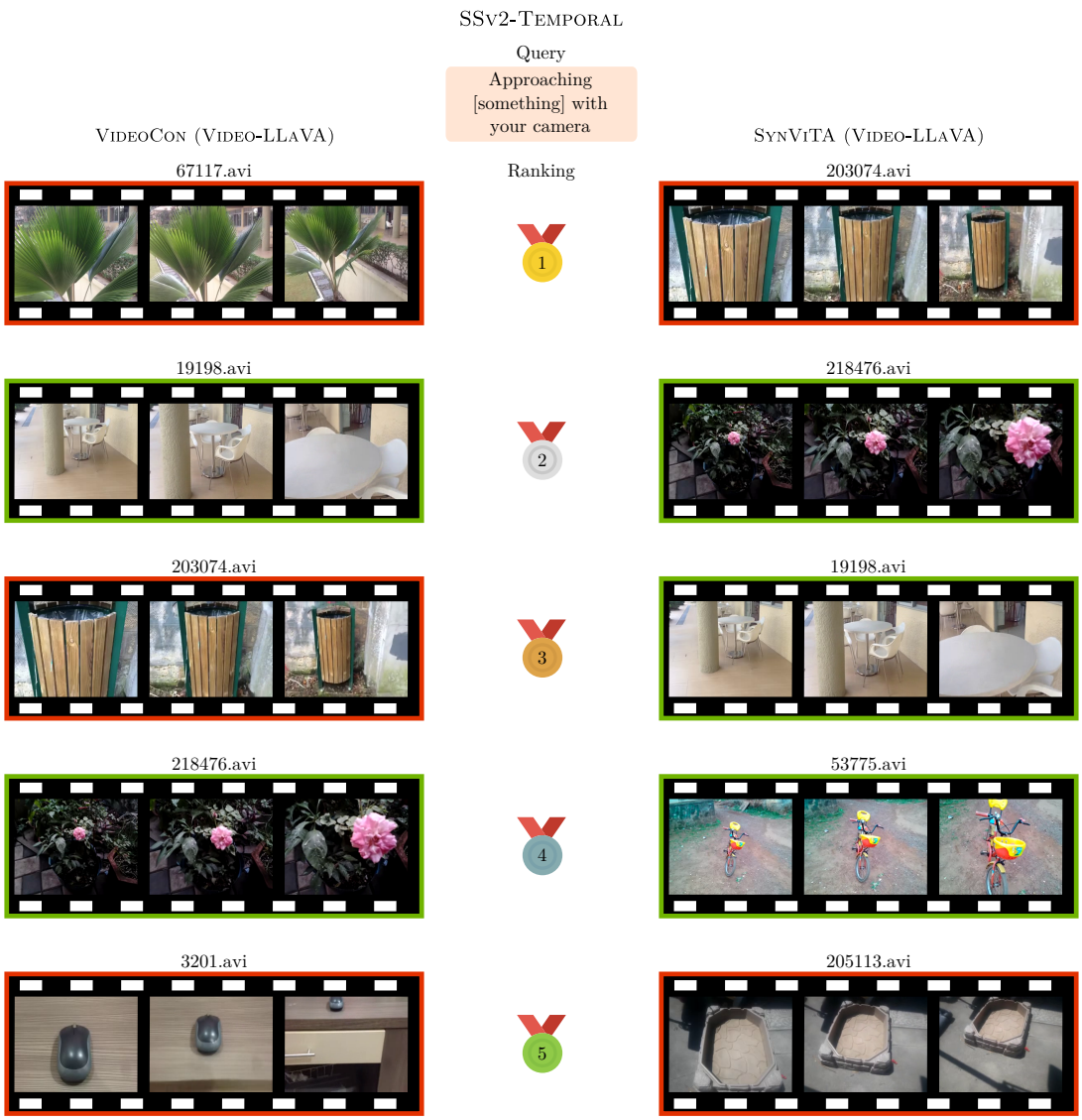

Figure 10. Comparison of rankings based on video-language alignment scores for the text-to-video retrieval task on SSv2-Temporal, using SynViTA (Video-LLaVA) against the baseline VideoCon (Video-LLaVA) trained without synthetic videos. Videos marked with green borders correctly match the input text query, while those marked with red borders do not.

Figure 11. Comparison of rankings based on video-language alignment scores for the text-to-video retrieval task on SSv2-Events, using SynViTA (mPLUG-Owl 7B) against the baseline VideoCon (mPLUG-Owl 7B) trained without synthetic videos. Videos marked with green borders correctly match the input text query, while those marked with red borders do not.

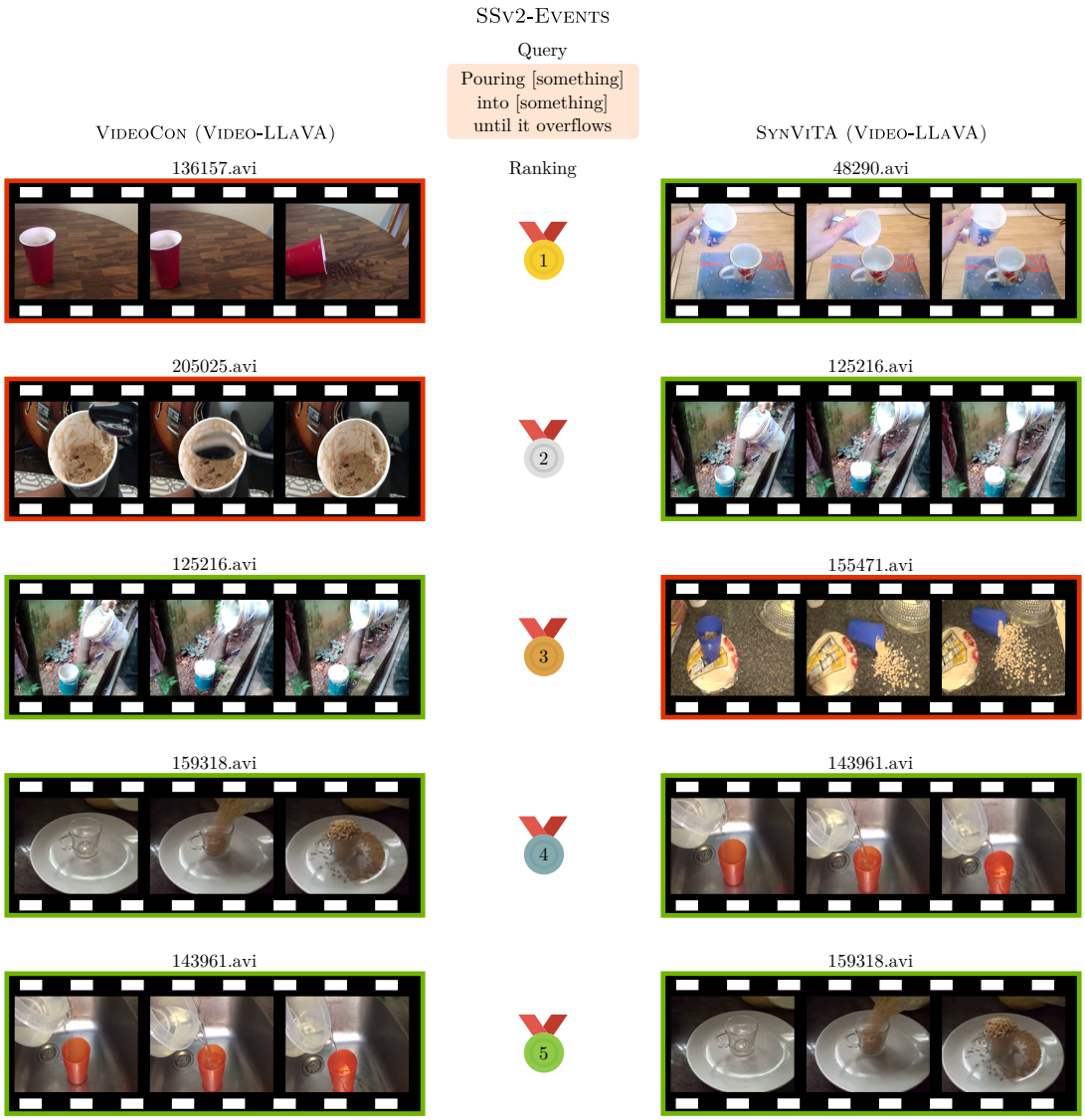

Figure 12. Comparison of rankings based on video-language alignment scores for the text-to-video retrieval task on SSv2-Events, using SynViTA (Video-LLaVA) against the baseline VideoCon (Video-LLaVA) trained without synthetic videos. Videos marked with green borders correctly match the input text query, while those marked with red borders do not.