SRL-CLIP: Efficient CLIP Video Adaptation via Structured Semantic Role Labels

Supplementary Material

Table 1. Impact of varying the no. of verb-role hard negatives, $\mathcal{N}_{vr}$.

	VidSitu	
$\mathcal{N}_{vr}$	Vb@1	CIDEr
0	52.38	75.63
1	51.72	74.78
2	52.54	74.02
3	53.26	74.88
4	**52.36**	**76.24**

Table 2. Impact of adapting different weights using LoRA. q, k, v, and o are the query, key, value, and output projection matrices in the self-attention block. fc and proj are the two MLPs after the self-attention module.

	VidSitu	
Weight type	Vb@1	CIDEr
q k v	**52.36**	**76.24**
q k v o	**52.62**	73.13
q k v o fc	51.22	73.61
q k v o fc proj	51.83	71.86

Appendix

We present additional results and discussions in the supplementary material. Sec. 1 starts with additional ablation experiments, focusing on hard negatives (HNs) and the low-rank adaptation (LoRA) module. In Sec. 2 we present qualitative results on the various tasks shown in the main paper. This provides an opportunity to better understand the working of SRL-CLIP. All ablation studies are conducted using the `ViT-L/14` variant.

1. Additional Ablations

How many Hard Negatives to use? Tab. 1 shows the impact of varying the number of hard negatives (HNs). Here, we see that as $\mathcal{N}_{vr}$ increases, the verb prediction accuracy increases while CIDEr drops by a small amount. This is expected as verb-role HNs tend to improve verb prediction performance. The best performance is considered at the geometric mean between Vb@1 and CIDEr on the VidSitu task.

What is the best LoRA configuration? In Tab. 2, we study the impact of adapting different weights of the SRL-CLIP image encoder with LoRA. We get the best performance when we include LoRA modules only for the attention weights in the Transformer ("q k v" – W_q, W_k, W_v) while keeping everything else frozen. Hence, in our default model, we only adapt the self-attention matrices (parameters) with LoRA.

Table 3. ZS T2V on MSRVTT with CLIP ViT B/32. We increase the amount of VidSitu training data, resulting in a consistent improvement in performance for R@10 and MnR.

	VidSitu	
Data	Vb@1	CIDER
0%	46.57	60.79
50%	50.82	70.72
70%	50.62	73.31
90%	51.36	73.55
100%	**52.36**	**76.24**

Training on subsets of VidSitu. The results of increasing the post-pretraining dataset from 10% to 100% are presented in Tab. 3. The table shows a consistent decline in mean rank and an improvement in R@10 with additional data. Additionally, with just 10% of the data, a mere 2,300 videos, we see a 3.9% boost in R@5 and a 2.6% in R@10; indicating the effectiveness of dense SRL prompts.

Detailed descriptions incur labeling costs, but save on compute costs. The use of dense prompts results in efficient training that completes in 5 hours on a *single* RTX2080 GPU (12 GB). In contrast, large models adapted on large (noisy) datasets incur significantly higher compute costs requiring many and larger GPUs. *E.g*. BT-Adapter [2] uses 8 V100 GPUs, TVTSv2 [3] uses 80 V100 GPUs, and VAST [1] uses 64 V100 GPUs. Considering the typical number of hyperparameter evaluations and experiments required, improving data quality is not only cheaper but also more viable in the long-term.

More examples of *natural* and *artificial* hard negatives are shown in Tab. 4. Notice how the naturally occurring hard negatives are good enough to learn strong video representations even without the need for artificial hard negatives. Different from most works that only use text-based negatives, VidSitu also facilitates visual natural negatives for the same text. Note how the hard negative captions are very plausible; in example 1 the action *look* instead of *speak*.

2. Qualitative Results

We now present qualitative results on 6 datasets. When not mentioned otherwise, we use the default variant of SRL-CLIP.

Positive Prompt	Natural Hard Negatives	Verb-role Hard Negatives
In this photo, the *action* is speak where, the *talker* is man standing in yellow sweatshirt, the *hearer* is woman with scarf, the *manner* is standing in the middle of a full airplane, the *scene* of the event is an airplane.	In this photo, the *action* is turn where, the *the turner* is man standing in yellow sweatshirt, the *the thing turning* is his body, the *direction* is towards woman with scarf, the *scene* of the event is an airplane.	In this photo, the *action* is look where, the *looker* is man standing in yellow sweatshirt, the *thing looked at* is woman with scarf, the *direction* is *is to his back*, *manner* is standing in the middle of a full airplane, the *scene* of the event is an airplane.
In this photo, the *action* is open where, the *opener* is man in brown jacket and man in gray suit, the *the thing opening* is trunk of taxi, the *manner* is annoyed, the *scene* of the event is near a taxi.	In this photo, the *action* is hoist where, the *lifter* is man in brown jacket and man in gray suit, the *thing going up* is dead body, the *direction* is up into trunk of taxi, the *scene* of the event is near a taxi.	In this photo, the *action* is respond where, the *replier* is man in brown jacket and man in gray suit, the *scene* of the event is near a taxi.
In this photo, the *action* is bow where, the *bower* is the woman in glasses, the *bowed to* is man wearing black, the *manner* is on her knees, the *scene* of the event is in a well lit room.	In this photo, the *action* is photograph, take a picture where, the *photographer* is the man in black suit, the *subject* is woman in glasses, *scene* of the event is in a well lit room.	In this photo, the *action* is smash where, the *smasher* is the woman in glasses, the *smashed* is man wearing black, the *direction* is on patients face, the *scene* of the event is in a well lit room.

Table 4. We show the naturally occurring hard negatives in a batch as well as the process of converting a standard positive prompt into hard negatives by swapping verb-role information. The template is shown in gray, *e.g.* In this photo,. The action and roles are shown in *italics*, *e.g. action, talker, hearer*. The correct prompt values (verbs or nouns) are in cobalt blue, *e.g. speak, man standing in yellow sweatshirt*; and the replaced verbs, roles, or nouns are in deep red. We swap the verb and roles in verb-role hard negatives while keeping the same nouns and performing some mapping between previous and new roles.

MSRVTT. We show zero-shot text-to-video retrieval on the MSRVTT dataset in Fig. 1. We can see that SRL-CLIP performs much better than CLIP when the queries have a compositional nature. The last row shows a failure case. Although SRL-CLIP retrieves a video in which a man is talking, he is not talking about hiking. It is hard to pick the right video just by using the visual modality, as hiking is not very clear by just watching the video. In fact, SRL-CLIP retrieves a video shot outdoors, which may have be associated with hiking rather than the indoor video.

LSMDC. We show zero-shot text-to-video retrieval on the LSMDC dataset in Fig. 2. LSMDC is a much harder dataset compared to MSRVTT, as it is based on movies that contain more dynamic shot changes. Also, the agent/patient of an action is annotated as "SOMEONE", unlike VidSitu, where they are described according to their characteristics, making it even more challenging. We can see that SRL-CLIP outperforms CLIP here as well.

VidSitu. Fig. 3 shows the qualitative results on video situation recognition for 5 videos. SRL-CLIP outperforms CLIP, especially when picking attributes like color. SRL-CLIP is also better at predicting the role *manner* (which captures the expression/emotion of the person), which CLIP struggles with. However, both SRL-CLIP and CLIP show similar (good) performance when predicting the *scene*. The last row shows a failure case (note that CLIP also fails to give good noun captions in this case). It is interesting to see that SRL-CLIP correctly identifies the *reacher* as a boy but assigns the wrong attribute to it.

References

[1] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[2] Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H Li, and Ge Li. One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*, 2023. 1

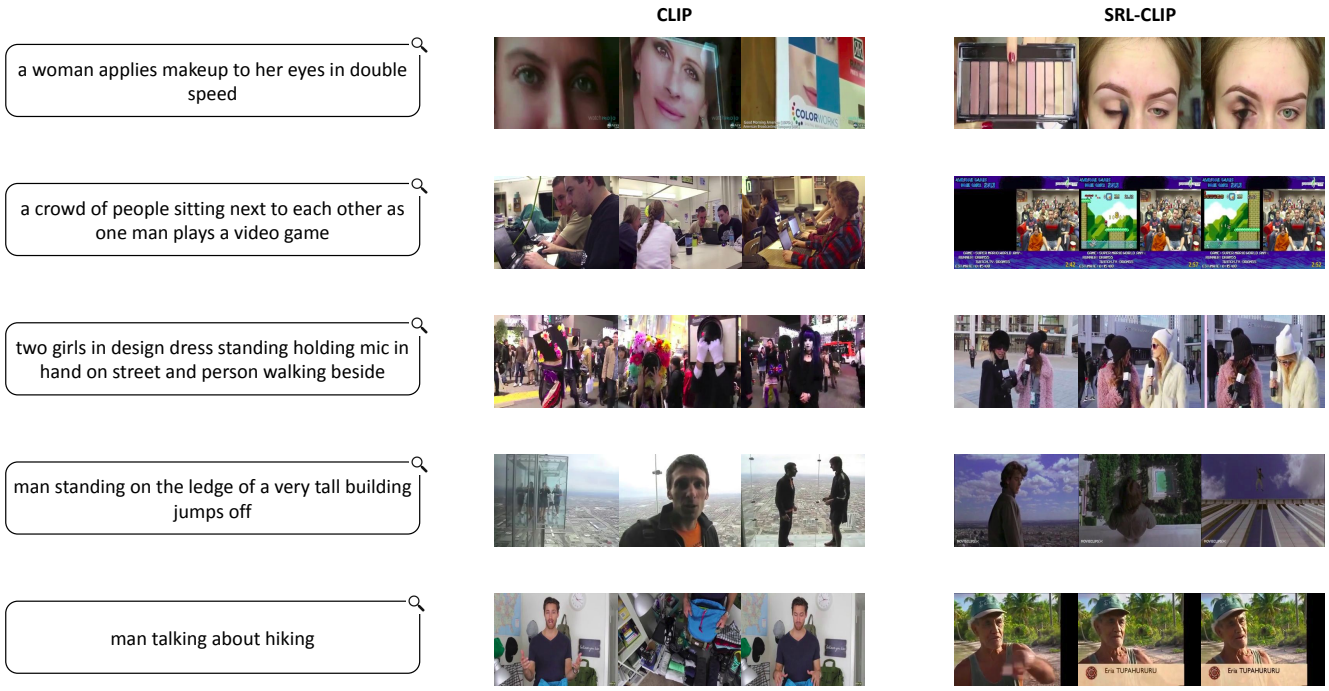

Figure 1. Zero-shot text-to-video retrieval on the MSRVTT dataset. We show three frames of the top-1 retrieved video for each query. We can see that SRL-CLIP outperforms CLIP, specially when compositional reasoning is required. The last row shows a failure case. Although SRL-CLIP retrieves a video in which a man is talking, and potentially with more appropriate background, he is not talking about hiking.

Figure 2. Zero-shot text-to-video retrieval on the LSMDC dataset. We show three frames of the top-1 retrieved video for each query. We can again notice that SRL-CLIP performs better than CLIP when compositional reasoning is needed. The last row shows a failure case.

[3] Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shu-Tao Xia, and Ying Shan. TVTSv2: Learning Out-of-the-box Spatiotemporal Visual Representations at Scale. *arXiv preprint arXiv:2305.14173*, 2023. 1

	verb	driver	vehicle	direction	manner	scene
CLIP	drive	man in hat	car	down the street	slowly	in a car
SRL-CLIP	drive	man in black jacket	car	down the road	with a serious look on his face	in a car
GT	drive	man in a black police uniform	car	forward	intently	car

	verb	entity entering	thing entered	manner	scene
CLIP	walk	man in blue shirt	the house	with his right hand	outside a house
SRL-CLIP	enter	boy in red shirt	the door	slowly	outside a building
GT	enter	kid in red	door	casually	in doorway

	verb	looker	looked at	direction	manner	scene
CLIP	stare	woman with dark hair	man in green shirt	down	with a slight smile	in a room
SRL-CLIP	stare	woman with blonde hair	man in brown jacket	down	with a sad expression	in a room
GT	look	girl with blonde hair	a man in front of her	forward	sadly	in a room

	verb	talker	hearer	manner	scene
CLIP	look	girl with dark hair	man in blue shirt	while seated next to each other	in a room
SRL-CLIP	speak	woman in a green shirt	man in brown shirt	while face to face	in a kitchen
GT	speak	blonde girl	old man	while sitting down	cabin

	verb	reacher	body part	goal	direction	purpose	scene
CLIP	kneel	woman in blue coat	hand	to grab something	down	to get something	in a room
SRL-CLIP	kneel	boy in orange shirt	his body	to grab something	down	to pick up a plate	in a room
GT	grab	boy	his hand	booklet	towards the man	to take booklet from man	lab

Figure 3. Video Situation Recognition on 5 videos. SRL-CLIP performs much better than CLIP in picking the right attribute of an entity. The last row shows a failure case where the semantic role labels predicted by SRL-CLIP deviates from the ground-truth (GT).