

# Appendix

## Table of Contents

<b>A Quantitative results</b>	<b>1</b>
A.1. Standard vs. Verb-Focus Contrastive (VFC) learning for all benchmarks . . . . .	1
A.2. MSR-VTT retrieval . . . . .	1
A.3. Additional ablations . . . . .	2
<b>B Qualitative results</b>	<b>4</b>
B.1. MSR-VTT . . . . .	4
B.2. NEXT-QA . . . . .	6
B.3. Kinetics-verb: Further analysis of calibration . . . . .	6
B.4. PaLM vs. rule-based methods for hard negative generation . . . . .	6
B.5. PaLM vs. rule-based methods for verb phrase extraction . . . . .	6
<b>C Baselines &amp; Implementation details</b>	<b>7</b>
C.1. Baselines . . . . .	7
C.2. CLIP4CLIP Architecture . . . . .	9
C.3. Fine-tuning details . . . . .	9
C.4. Evaluation protocols . . . . .	10
C.5. PaLM prompting . . . . .	10
C.6. T5 generations . . . . .	13
C.7. Kinetics-verb . . . . .	13

This appendix to the main paper provides additional quantitative (Sec. A) and qualitative results (Sec. B), and further details on baselines and implementation (Sec. C).

Method	loss	MSR-VTT		K-400		NEXT-QA		SVO	
		3k val. MC	Verb <sub>H</sub> MC	all top-1	verb top-1	all AP	ATP <sub>hard</sub> MC	all AP	verb AP
ZERO-SHOT									
Baseline	NCE	94.9	69.9	55.6	52.1	48.6	28.9	60.2	61.9
VFC	NCE	94.9	78.3	58.5	56.7	51.0	31.3	61.5	63.9
VFC	HardNeg-NCE	<b>95.1</b>	<b>80.5</b>	<b>58.8</b>	<b>57.1</b>	<b>51.5</b>	<b>31.4</b>	<b>61.8</b>	<b>64.6</b>
FINED-TUNED									
Baseline	NCE	<b>96.8</b>	73.8	-	-	57.3	37.8	-	-
VFC	NCE	96.2	84.8	-	-	58.4	38.3	-	-
VFC	HardNeg-NCE	96.2	<b>85.2</b>	-	-	<b>58.6</b>	<b>39.3</b>	-	-

Table A.1. **Standard vs. Verb-Focus Contrastive learning for all benchmarks.** We report MSR-VTT random (3k val.) and Verb<sub>H</sub> [53] multiple-choice accuracies, Kinetics-400 and Kinetics-verb top-1 accuracies, NEXT-QA and ATP<sub>hard</sub> [9] multiple-choice accuracies, and SVO-probes entire dataset and verb-focused Average Precision. We observe that our VFC learning performs better than standard contrastive learning (Baseline) for all verb-focused benchmarks on both zero-shot and fine-tuned settings, while maintaining performance on more noun-focused benchmarks, such as MSR-VTT random MC. We observe that using the HardNeg-NCE loss, instead of standard NCE, further improves performance for all benchmarks on both zero-shot and fine-tuned settings.

## A. Quantitative results

In this section, we present results comparing standard versus Verb-Focused Contrastive (VFC) learning for all benchmarks (Sec. A.1), comparison to state-of-the-art methods for MSR-VTT retrieval (Sec. A.2), and additional ablations (Sec. A.3).

### A.1. Standard vs. Verb-Focus Contrastive (VFC) learning for all benchmarks

We see in Tab. A.1 that our VFC learning performs better than standard contrastive learning (Baseline) for all verb-focused benchmarks on both zero-shot and fine-tuned settings while maintaining performance on more noun-focused benchmarks, such as MSR-VTT random MC. We observe that using the HardNeg-NCE loss, instead of standard NCE, further improves performance for all benchmarks on both zero-shot and fine-tuned settings.

### A.2. MSR-VTT retrieval

We see in Tab. A.2 that while our verb-focused pre-training drastically improves performance on verb-focused benchmarks – such as Verb<sub>H</sub> split [53] MSR-VTT (see main paper Tab. 9) – it maintains performance on noun-focused benchmarks such as MSR-VTT retrieval T2V (1k split) in a zero-shot setting. We perform comparably to InternVideo [75] in a zero-shot setting, while using a significantly smaller setting both in terms of architecture (InternVideo uses 2.8× more parameters and 12.4× more flops) and pretraining dataset size (they use 24× more data). In a fine-tuned setting, InternVideo surpasses VFC’s perfor-

Model	# params.	1K val. T→V R@1
ZERO-SHOT		
VideoCLIP [84]	–	10.4
CLIP [58]	151M	30.6
InternVideo [75]‡	≈ 460M	<b>40.7</b>
VFC (Ours)	164M	40.3
FINED-TUNED		
ClipBERT [38]	–	22.0
MMT [26]	–	26.6
VideoCLIP [84]	–	30.9
CLIP-straight [55]	151M	31.2
MMT (CLIP features) [26]	–	34.0
C4CL-mP [53]	151M	43.1
CLIP2Video [53]	–	45.6
InternVideo [75]‡	≈ 460M	<b>55.2</b>
VFC (Ours)	164M	44.5

Table A.2. **Results on MSR-VTT retrieval.** We report T2V retrieval on the 1k split. While our VFC framework drastically improves performance on verb-focused benchmarks, including Verb<sub>H</sub> split [53] (see main paper Tab. 9), it maintains performance on noun-focused benchmarks such as the retrieval 1k split in the zero-shot setting. In the fine-tuned setting, InternVideo surpasses VFC’s performance. ‡ InternVideo is concurrent unpublished work with a larger model (2.8× more parameters and 12.4× more flops), and has a larger pretraining dataset size (they use 24× more data).

mance. This is expected given our model parameters and flops are significantly smaller – see number of parameters in Tab. A.2.

### A.3. Additional ablations

Here, we present ablation results for video mining (Sec. A.3.1), PaLM prompting (Sec. A.3.2), the verb phrase loss (Sec. A.3.3), fine-tuning strategy (Sec. A.3.4), calibration (Sec. A.3.5), hard noun vs. verb negatives (Sec. A.3.6), and training with hard negatives for other datasets (Sec. A.3.7). We note that all ablations are performed with the standard NCE loss (not HardNeg-NCE).

#### A.3.1 Video mining

An alternative to our proposed calibration strategy to avoid imbalances due to the addition of negative captions would be to avoid training with unpaired data at all, by mining a matching video  $V_{i_k}^{\text{hard}}$  for each generated caption  $T_{i_k}^{\text{hard}}$ . We attempt this via CLIP-based text-to-video retrieval in a large video database. We next explain our pipeline in more detail.

Firstly, we generate hard negative verb captions with PaLM as explained in Sec. 3.2 of the main paper. For each hard negative caption, we then perform text-to-image retrieval to find a matching video in the VideoCC [50]

Method	# pairs	Verb <sub>H</sub>	K-400	SMiT
Baseline	481K	69.9	55.6	78.3
HN	481K	78.0 (+8.1)	55.8 (+0.2)	78.6 (+0.3)
HN+VM	1.22M	78.7 (+8.8)	51.8 (-3.8)	75.0 (-3.3)

Table A.3. **Video Mining.** We report multi-choice accuracy on Verb<sub>H</sub> [53], Kinetics-400 top-1 accuracy and V2T R@1 on Spoken Moment in Time (validation set of our pretraining SMiT data). We observe that although our Video Mining (VM) approach improves performance on Verb<sub>H</sub>, it causes a drop in performance on Kinetics and SMiT, which highlights that the additional video-text pairs are noisy. For experiments including hard negatives, we note that one hard negative is sampled for each video here.

database. Specifically, we calculate the cosine similarity between the hard negative caption CLIP text embedding and the average of the video frames’ CLIP image embedding, for all videos in the database. We then keep the video with closest similarity to the hard negative caption to form a new video-text pair. Finally, we apply a similarity threshold to keep only the best matching video-text pairs and add these to our training set. In practice, we experiment with different thresholds and find a value of 0.28 to work best, adding a total of 738K new video-text pairs to training (SMiT training set size is 481K). Note that we also experiment with text-to-text retrieval: in this case, we calculate the similarity between each hard negative caption and all VideoCC captions, and subsequently use the video corresponding to the closest VideoCC caption to form a new pair. However, we find this performs worse.

We observe in Tab. A.3 that our additional video-text pairs are noisy. In fact, although this approach improves performance on Verb<sub>H</sub>, it causes a large drop in performance on Kinetics and SMiT (validation set of our pretraining data). Finding a video matching a specific, detailed and long caption is challenging (see qualitative examples in Fig. A.1). A video matching the caption may not exist in the VideoCC corpus and even if it did, for this method to be successful, the mined video must match the generated caption on the verbs (and CLIP is biased towards images and objects only, which is exactly the problem we are trying to solve).

#### A.3.2 Giving input-output example pairs to PaLM

To generate hard verb negative captions with PaLM, we also add four input-output pair examples to the prompt (see full prompt in Sec. C.5) to increase the quality of the generated hard negatives. We observe in Tab. A.4 that the input-output pairs improve the performance on Verb<sub>H</sub> and Kinetics-400.

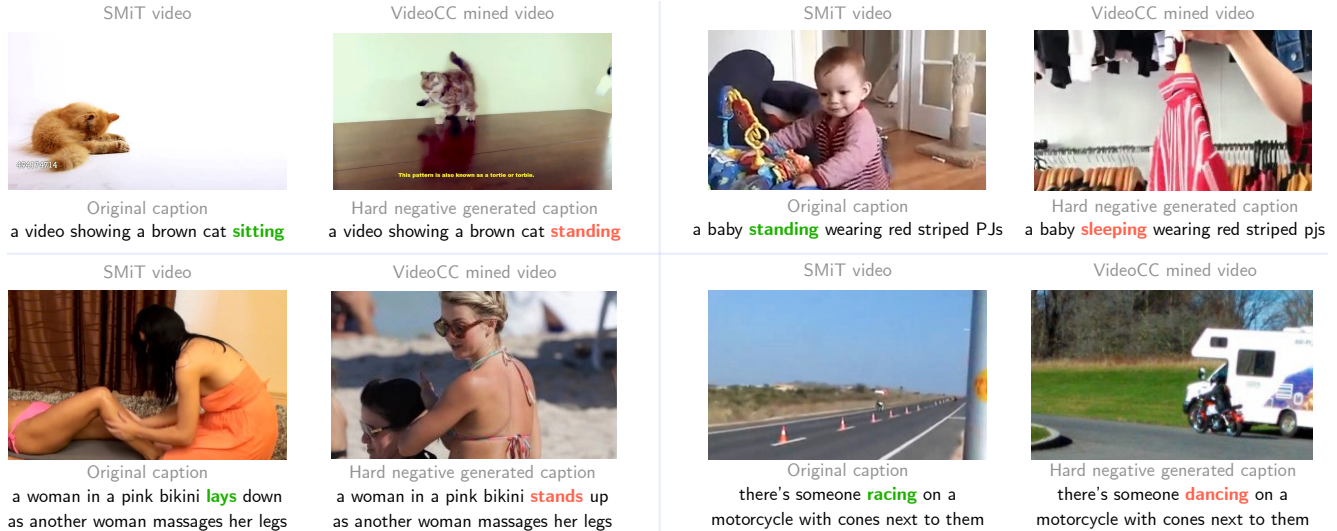


Figure A.1. **Video mining:** We show examples of mined matching videos for generated hard negative captions. For ease of visualisation, we show a single frame per video. In some cases, as the top left corner, the mined video from VideoCC closely matches the hard negative caption. However, often, the new video-text pairs are noisy. For example, in the top right corner, the mined video contains a ‘red striped shirt’ but no ‘baby sleeping’. In the bottom left example, there is a ‘woman in a pink bikini standing up’ but no ‘woman massaging her legs’. Finally, in the bottom right example, although the video contains a ‘motorcycle’, the person is not ‘dancing’ and there are no ‘cones next to them’.

input-output pairs	Verb <sub>H</sub>	K-400
✓	77.5 78.0	54.6 55.8

Table A.4. **Inclusion of input-output pairs in PaLM prompt.** We report multi-choice accuracy on Verb<sub>H</sub> [53] and Kinetics-400 top-1 accuracy. We observe that including input-output pairs in the PaLM prompt for generating hard negative captions increases the performance on both benchmarks. We note that one hard negative is sampled for each video here.

### A.3.3 Verb phrase loss

We see in Tab. A.5 that using only the video-to-text component of the verb phrase loss allows us to maintain performance on noun-focused benchmarks such as MSR-VTT retrieval, while also giving a performance boost on verb focused benchmarks Verb<sub>H</sub> and K-400.

### A.3.4 Fine-tuning image and text towers

We experiment with different fine-tuning strategies: (i) fine-tuning both image and text towers, (ii) freezing the image CLIP backbone only (here, the sequence Transformer seqTrans and text tower are trained – see Sec. C.2 for more details on the CLIP4CLIP architecture), (iii) freezing the text tower only (here, seqTrans and image CLIP backbone are trained), (iv) freezing both image and text towers (here, only seqTrans is trained). We see in Tab. A.6 that fine-

Method	Verb phrase loss		MSR-VTT		K-400
	T→V	V→T	1k val. T→V R@1	Verb <sub>H</sub> MC acc	all Top-1
Baseline			40.8	69.9	55.6
VFC (Ours)	✓	✓	38.8 (-2.0)	77.0 (+7.1)	58.8 (+3.2)
VFC (Ours)		✓	40.1 (-0.7)	76.3 (+6.4)	58.5 (+2.9)

Table A.5. **Verb phrase loss.** We report MSR-VTT T2V retrieval on the 1k split, multi-choice accuracy on Verb<sub>H</sub> [53] and Kinetics-400 top-1 accuracy. We observe that using only the video-to-text component of the verb phrase loss allows us to maintain performance on noun-focused benchmarks such as MSR-VTT retrieval, while also giving a performance boost on Verb<sub>H</sub> and K-400. For experiments including hard negatives, we note that one hard negative is sampled for each video here.

tuning both image and text towers works best. We do not include setting (iv) as it performs very poorly.

### A.3.5 Calibration

As explained in Sec. 3.2 of the main paper, our calibration strategy is composed of two steps: (1) ignoring hard negative captions from the other elements of the batch (denoted as ‘reducing  $B$  effect’, where  $B$  is the batch size); (2) filtering the generated PaLM captions to have equal number of concept occurrences in positive and negative pairs (denoted as  $G_{\omega} \approx S_{\omega}$ ). We show the effect of each of these steps in Tab. A.7. We observe that by combining both steps, we avoid a drop in performance on Kinetics-400, while main-

Method	✱text	✱image	Verb <sub>H</sub>	K-400
Baseline			69.9	55.6
VFC (Ours)			76.3	58.5
VFC (Ours)	✓		72.0 (-4.3)	54.8 (-3.7)
VFC (Ours)		✓	75.1 (-1.2)	55.1 (-3.4)

Table A.6. **Fine-tuning image and text towers.** We report multi-choice accuracy on Verb<sub>H</sub> [53] and Kinetics-400 top-1 accuracy. ✱corresponds to freezing the image or text tower. We observe that fine-tuning both image and text towers works best. For experiments including hard negatives, we note that one hard negative is sampled for each video here.

Method	reducing $B$ effect	$G_\omega \approx S_\omega$	Verb <sub>H</sub>	K-400
Baseline			69.9	55.6
HN			80.5	54.5
HN	✓		79.4	55.4
HN		✓	78.7	55.2
HN	✓	✓	78.0	55.8

Table A.7. **Calibration strategy.** We report multi-choice accuracy on Verb<sub>H</sub> [53] and Kinetics-400 top-1 accuracy. We observe that by combining both calibration steps, we avoid a drop in performance on Kinetics-400, while maintaining a large performance improvement on Verb<sub>H</sub>. For experiments including hard negatives, we note that one hard negative is sampled for each video here.

taining a large performance improvement on Verb<sub>H</sub>.

### A.3.6 Hard noun vs. verb negatives

We modify our prompt to task PaLM to generate hard noun negatives, and compare to training with verb negatives in Tab. A.8. We find that noun negative training only slightly outperforms the baseline on Verb<sub>H</sub> (71.0 vs. 69.9). This makes sense since Verb<sub>H</sub> is a verb-focused benchmark. We also evaluate on SVO-Probes: we find that training with hard verb negatives performs best on the verb split, while training with noun negatives performs best on the object split. We leave exploring hard noun negatives for future work.

HN	Verb <sub>H</sub>	SVO <sub>verb</sub>	SVO <sub>object</sub>
∅	69.9	61.9	79.5
PaLM verb	<b>78.0</b>	<b>62.9</b>	79.1
PaLM noun	71.0	61.6	<b>80.0</b>

Table A.8. **PaLM verb vs. noun hard negative (HN) training.** On Verb<sub>H</sub>, noun negative training only slightly outperforms the baseline which makes sense since this is a verb-focused benchmark. On SVO-Probes, we find that training with hard verb negatives performs best on the verb split, while training with noun negatives performs best on the object split.

### A.3.7 Hard verb negative training with other datasets

We generate hard verb negatives using PaLM for two additional datasets – HowTo1M (1M subset of HowTo100M), and MSR-VTT directly (Tab. A.9). For Howto1M, we observe that (i) pretraining is less effective than SMiT and (ii) hard negative training does not bring gains. We hypothesise this is due to the noisy captions: they correspond to ASR outputs, are less aligned to the visual content, and sometimes correspond to one word sentences. However, for MSR-VTT, hard negative training gives a large boost on Verb<sub>H</sub> (82.1 vs. 74.8), showing that our method is not restricted to SMiT.

Dataset	w/o HN	w/ HN
SMiT	69.9	78.0 (+8.1)
HowTo1M	67.5	66.4 (-1.1)
MSR-VTT	74.8	82.1 (+7.3)

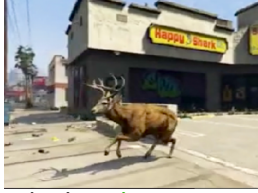
Table A.9. **Hard negatives (HN) with different datasets.** For Howto1M which has noisy captions, we observe that (i) pretraining is less effective than SMiT and (ii) hard negative training does not bring gains. However, for MSR-VTT, hard negative training gives a large boost on Verb<sub>H</sub> (82.1 vs. 74.8), showing that our method is not restricted to SMiT.

## B. Qualitative results

In this section, we present qualitative results on MSR-VTT (Sec. B.1) and NEXT-QA (Sec. B.2), further analysis of calibration on Kinetics-verb (Sec. B.3), and comparisons of the use of PaLM *versus* rule-based methods for hard negative (Sec. B.4) and verb phrase (Sec. B.5) generations.

### B.1. MSR-VTT

We show qualitative examples from the Verb<sub>H</sub> [53] multiple choice evaluation in Figure A.2. For each video sample, we show the 5 captions ranked in order of decreasing similarity for both our baseline and VFC models. We observe that the baseline model often mistakes the hard negative as matching the video. This effect is reduced when training with hard negatives, as proposed in our VFC method, enabling the correct caption to be retrieved from the 5 options. In some rare cases, as shown on the last row, the baseline model is correct but training with hard negatives causes the hard negative to have highest similarity with the video. For example, the model incorrectly ranks ‘a silent clip of a woman **smiling** at people’ higher than ‘a silent clip of a woman **screaming** at people’.



a deer is **running** across a road in a video game

**Baseline (Ours)**

1. a deer is **rolling** across a road in a video game
2. a deer is **running** across a road in a video game
3. kids are reacting to viral videos
4. a group of people celebrating some kind of festival
5. models are walking down a short runway

**VFC (Ours)**

1. a deer is **running** across a road in a video game
2. a deer is **rolling** across a road in a video game
3. kids are reacting to viral videos
4. a group of people celebrating some kind of festival
5. models are walking down a short runway



a cartoon girl **sings** about rain

**Baseline (Ours)**

1. a cartoon girl **dances** about in the rain
2. a cartoon girl **sings** about rain
3. a bird white color is dancing
4. a judge talks to a young performer on the stage
5. man using spary on the underside of a car

**VFC (Ours)**

1. a cartoon girl **sings** about rain
2. a cartoon girl **dances** about in the rain
3. a bird white color is dancing
4. a judge talks to a young performer on the stage
5. man using spary on the underside of a car



a basketball player **shoots** a layup

**Baseline (Ours)**

1. a basketball player **misses** a layup
2. a basketball player **shoots** a layup
3. this is a video advertisement about a fruit juice
4. a woman is talking
5. different couples are shown at a table

**VFC (Ours)**

1. a basketball player **shoots** a layup
2. a basketball player **misses** a layup
3. this is a video advertisement about a fruit juice
4. a woman is talking
5. different couples are shown at a table



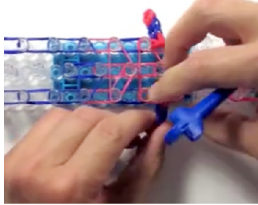
two wrestlers are **fighting** in the ring

**Baseline (Ours)**

1. two wrestlers are **hugging** each other in the ring
2. two wrestlers are **fighting** in the ring
3. two men are on a hill
4. four young girls are sitting and laughing
5. a person is quickly dicing up the onions

**VFC (Ours)**

1. two wrestlers are **fighting** in the ring
2. two wrestlers are **hugging** each other in the ring
3. two men are on a hill
4. four young girls are sitting and laughing
5. a person is quickly dicing up the onions



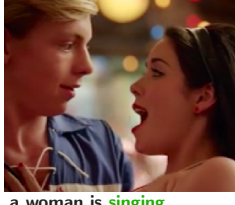
a man is **building** a vehicle

**Baseline (Ours)**

1. a man is **sketching** a vehicle
2. a man is **building** a vehicle
3. there is a commentary about the arriving of vips
4. a photoshop tutorial featuring a photo of a woman
5. there is a woman is singing a new song

**VFC (Ours)**

1. a man is **building** a vehicle
2. a man is **sketching** a vehicle
3. there is a commentary about the arriving of vips
4. a photoshop tutorial featuring a photo of a woman
5. there is a woman is singing a new song



a woman is **singing** while staring at a man

**Baseline (Ours)**

1. a woman is **crying** while staring at a man
2. a woman is **singing** while staring at a man
3. a beautiful video presentation of chef jon favreau
4. a little girl plays with a small play set
5. fanfare on an indoor soccer field is being shown

**VFC (Ours)**

1. a woman is **singing** while staring at a man
2. a woman is **crying** while staring at a man
3. a beautiful video presentation of chef jon favreau
4. a little girl plays with a small play set
5. fanfare on an indoor soccer field is being shown



a man is **giving** a political speech

**Baseline (Ours)**

1. a man is **listening** to a political speech
2. a man is **giving** a political speech
3. an interesting scene of wrestling
4. a lot of people walking around a small space
5. a sloth is climbing in a tree

**VFC (Ours)**

1. a man is **giving** a political speech
2. a man is **listening** to a political speech
3. an interesting scene of wrestling
4. a lot of people walking around a small space
5. a sloth is climbing in a tree



a person **standing** on the edge of a building

**Baseline (Ours)**

1. a person **sleeping** on the edge of a building
2. a person **standing** on the edge of a building
3. she routed on hilary
4. this is a video from the voice kids
5. a man performs a ping pong trick shot

**VFC (Ours)**

1. a person **standing** on the edge of a building
2. a person **sleeping** on the edge of a building
3. she routed on hilary
4. this is a video from the voice kids
5. a man performs a ping pong trick shot



a video of a guy **driving** a lamborghini

**Baseline (Ours)**

1. a video of a guy **crashing** a lamborghini
2. a video of a guy **driving** a lamborghini
3. a man says he is going to go to samokov to buy some stuff
4. korean guy singing
5. people gathered in a place and release balloons in the air

**VFC (Ours)**

1. a video of a guy **driving** a lamborghini
2. a video of a guy **crashing** a lamborghini
3. a man says he is going to go to samokov to buy some stuff
4. korean guy singing
5. people gathered in a place and release balloons in the air



chef is **cooking** food here

**Baseline (Ours)**

1. chef is **servng** food here
2. chef is **cooking** food here
3. a old man wears specs talking to media
4. an animation talking about economists
5. someone is showing how to solve a rubik cube

**VFC (Ours)**

1. chef is **cooking** food here
2. chef is **servng** food here
3. a old man wears specs talking to media
4. someone is showing how to solve a rubik cube
5. an animation talking about economists



a group of people **yelling** at the camera

**Baseline (Ours)**

1. a group of people **yelling** at the camera
2. a group of people **dancing** at the camera
3. this is a video of chinese guy rapping
4. a man has his arms crossed
5. a child is preparing to bake something

**VFC (Ours)**

1. a group of people **dancing** at the camera
2. a group of people **yelling** at the camera
3. this is a video of chinese guy rapping
4. a man has his arms crossed
5. a child is preparing to bake something



a silent clip of a woman **screaming** at people

**Baseline (Ours)**

1. a silent clip of a woman **screaming** at people
2. a silent clip of a woman **smiling** at people
3. texts are being sent back and forth
4. a man speaks about a teaching curriculum
5. the boy is trying to fix the problem

**VFC (Ours)**

1. a silent clip of a woman **smiling** at people
2. a silent clip of a woman **screaming** at people
3. texts are being sent back and forth
4. a man speaks about a teaching curriculum
5. the boy is trying to fix the problem

Figure A.2. MSR-VTT verb-focused benchmark: We show qualitative examples from the Verb<sub>H</sub> [53] multiple choice evaluation. For ease of visualisation, we only show a single frame per video. For each video sample, we show the 5 captions ranked in order of decreasing similarity for both our baseline and VFC models. We observe that the baseline model often mistakes the hard negative as matching the video; for example in the top left example, the caption ‘a deer is rolling across a road in a video game’ is ranked higher than the correct answer ‘a deer is running across a road in a video game’. When training with hard negatives as in our VFC model, the model performance improves, retrieving the correct caption from the 5 options. On the bottom row, we show two failure cases where training with hard negatives causes the model to make a mistake; choosing the hard negative (‘a group of people dancing at the camera’, ‘a silent clip of a woman smiling at people’) as the correct caption instead of (‘a group of people yelling at the camera’, ‘a silent clip of a woman screaming at people’).

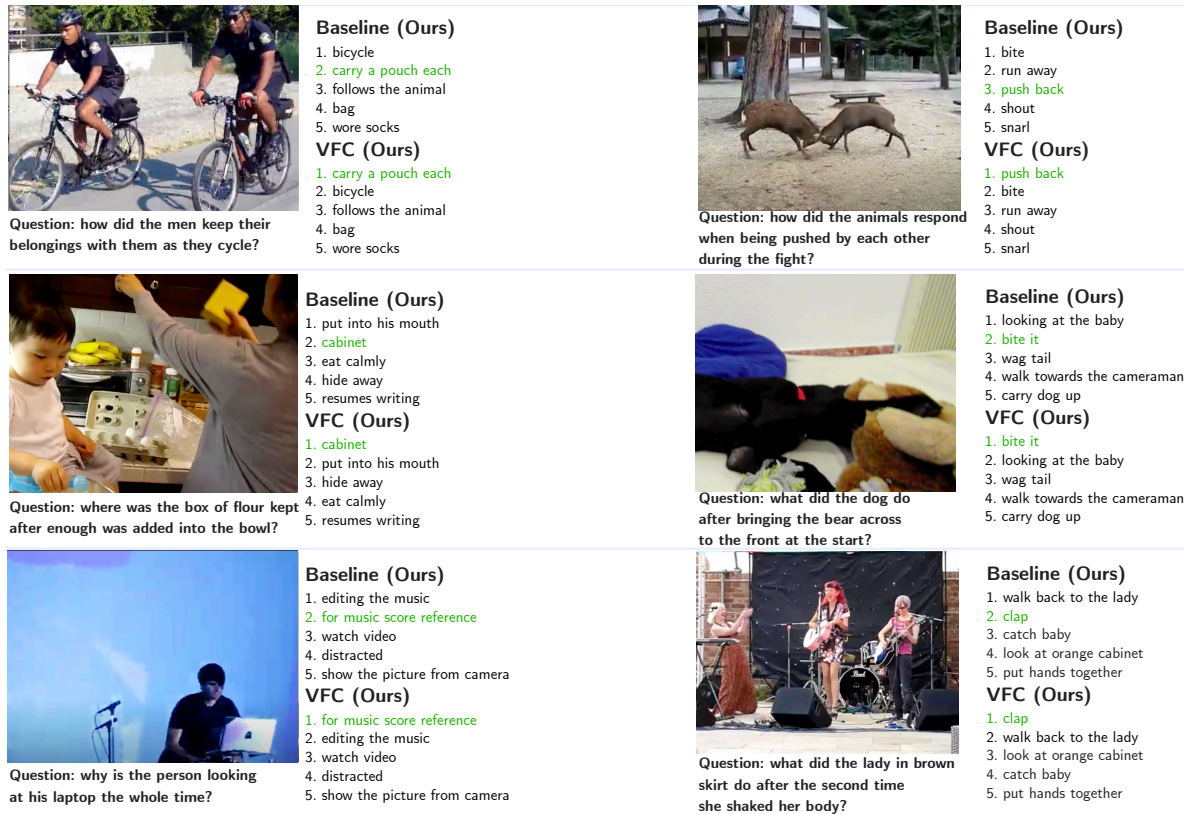


Figure A.3. **NEXT-QA**: We show qualitative examples from the  $ATP_{hard}$  [9] multiple choice evaluation. For ease of visualisation, we only show a single frame per video. For each video sample, we show the 5 answers ranked in order of decreasing similarity for both our baseline and VFC models in a zero-shot setting. We observe that our VFC model improves performance, retrieving the correct answer from the 5 options more often.

## B.2. NEXT-QA

In Figure A.3, we show qualitative examples from the  $ATP_{hard}$  [9] multiple choice evaluation. For each video sample, we show the 5 answers ranked in order of decreasing similarity for both our baseline and VFC models in a zero-shot setting. We observe that our VFC model improves performance, retrieving the correct answer from the 5 options more often.

## B.3. Kinetics-verb: Further analysis of calibration

In Tab. A.10, we show further examples of confusion matrices comparing training performance with *versus* without calibration on Kinetics-verb classes as in Tab. 4 of the main paper. Once again, we observe that calibration reduces the effect of ‘attraction’ points, which distort the feature space, by making  $R_{\omega}$  the same for all verb phrase concepts.

## B.4. PaLM vs. rule-based methods for hard negative generation

In Fig. A.4, we compare hard negative caption generation using PaLM to T5 and rule-based methods such as re-

placing detected verbs by random verbs or antonym verbs. We observe that LLM based methods result in linguistically and semantically viable sentences (which may not be guaranteed with random and antonym verb replacements). We also note that LLM based methods can change more than just the verb: (i) T5 and PaLM can replace the verb by a verb-noun pair and, (ii) PaLM can replace pronouns and determiners anywhere in the sentence (as opposed to T5 which can only replace the verb, see more details in Sec. C.6), making the negative caption more linguistically correct.

## B.5. PaLM vs. rule-based methods for verb phrase extraction

In Fig. A.5, we compare verb phrase extraction using PaLM to: (i) using action labels for clips from the Moments in Time (MiT) dataset (these are available as SMiT data inherits from MiT [48]) and (ii) using a rule-based method (NLTK [8]) to isolate verbs. We observe that using PaLM outperforms both: (i) MiT action labels can be general and conceal fine-grained action information in the video which can improve verb understanding, (ii) NLTK

has difficulties extracting all verbs in a sentence, and can often mistake them for nouns; NLTK also cannot extract a verb phrase when a verb is not present in the sentence (e.g. for the caption ‘this is an aerial shot of a very nice waterfall’, NLTK extracts no verb phrase while PaLM extracts ‘water flowing’); finally, our NLTK approach does not extract verb-noun pairs (e.g. for the caption ‘this is a video of two women who are doing gymnastics’, NLTK extracts ‘doing’ while PaLM extracts ‘doing gymnastics’) – this can be crucial for understanding the action in the video. We note that although NLTK could be used to extract verbs and nouns independently through PoS tagging, correctly assigning nouns to the matching verb is not always robust for long, complex sentences as in SMiT. Indeed, the average length of a sentence in SMiT is 18 words.

$R_\omega \propto$	w/o calibration	w/ calibration								
mopping floor	<table border="1"><tr><td>7</td><td>17</td></tr><tr><td>17</td><td>3</td></tr></table>	7	17	17	3	<table border="1"><tr><td>1</td><td>1</td></tr><tr><td>46</td><td>12</td></tr></table>	1	1	46	12
7	17									
17	3									
1	1									
46	12									
cleaning floor	<table border="1"><tr><td>17</td><td>15</td></tr><tr><td>17</td><td>3</td></tr></table>	17	15	17	3	<table border="1"><tr><td>12</td><td>36</td></tr><tr><td>12</td><td>36</td></tr></table>	12	36	12	36
17	15									
17	3									
12	36									
12	36									
$R_\omega \propto$										
dunking basketball	<table border="1"><tr><td>4</td><td>17</td></tr><tr><td>42</td><td>5</td></tr></table>	4	17	42	5	<table border="1"><tr><td>1</td><td>1</td></tr><tr><td>25</td><td>20</td></tr></table>	1	1	25	20
4	17									
42	5									
1	1									
25	20									
shooting basketball	<table border="1"><tr><td>29</td><td>8</td></tr><tr><td>29</td><td>8</td></tr></table>	29	8	29	8	<table border="1"><tr><td>7</td><td>39</td></tr><tr><td>7</td><td>39</td></tr></table>	7	39	7	39
29	8									
29	8									
7	39									
7	39									
$R_\omega \propto$										
doing nails	<table border="1"><tr><td>14</td><td>46</td></tr><tr><td>42</td><td>6</td></tr></table>	14	46	42	6	<table border="1"><tr><td>1</td><td>1</td></tr><tr><td>43</td><td>11</td></tr></table>	1	1	43	11
14	46									
42	6									
1	1									
43	11									
cutting nails	<table border="1"><tr><td>8</td><td>6</td></tr><tr><td>8</td><td>6</td></tr></table>	8	6	8	6	<table border="1"><tr><td>5</td><td>19</td></tr><tr><td>5</td><td>19</td></tr></table>	5	19	5	19
8	6									
8	6									
5	19									
5	19									

Table A.10. **Confusion matrix for Kinetics-verb classes.** Without proper calibration, the verb phrases ‘mopping floor’, ‘dunking basketball’, ‘doing nails’ become highly attractive in the video-text feature space. Our calibration mechanism alleviates this issue by making the ratio  $R_\omega$  independent of verb phrases (see details in Sec. 3.2 of the main paper).

## C. Baselines & Implementation details

In this section, we present detailed descriptions of baselines (Sec. C.1), the CLIP4CLIP [44] architecture used in all our experiments (Sec. C.2), fine-tuning (Sec. C.3) and evaluation protocols (Sec. C.4), the PaLM prompting pro-





<p>Original caption: a woman <b>squats</b> with an empty bar that has a couple of rubber bands attached to it on the floor</p> 	<p>Hard negative generated captions:</p> <p>Random verb: a woman <b>gardens</b> with an empty bar that has a couple of rubber bands attached to it on the floor</p> <p>Antonym verb: <b>[no antonym]</b></p> <p>T5: a woman <b>walks</b> with an empty bar that has a couple of rubber bands attached to it on the floor</p> <p>PaLM: a woman <b>runs to</b> an empty bar that has a couple of rubber bands attached to it on the floor</p>
<p>Original caption: people are <b>walking</b> around the mall that is somewhat crowded</p> 	<p>Hard negative generated captions:</p> <p>Random verb: people are <b>encouraging</b> around the mall that is somewhat crowded</p> <p>Antonym verb: people are <b>riding</b> around the mall that is somewhat crowded</p> <p>T5: people are <b>sitting</b> around the mall that is somewhat crowded</p> <p>PaLM: people are <b>driving</b> around the mall that is somewhat crowded</p>
<p>Original caption: a man is <b>sitting</b> on his bike on his cell phone</p> 	<p>Hard negative generated captions:</p> <p>Random verb: a man is <b>wrestling</b> on his bike on his cell phone</p> <p>Antonym verb: a man is <b>standing</b> on his bike on his cell phone</p> <p>T5: a man is <b>standing with a woman</b> on his bike on his cell phone</p> <p>PaLM: a man is <b>riding</b> his bike on his cell phone</p>
<p>Original caption: video of a man <b>texting</b> on his phone</p> 	<p>Hard negative generated captions:</p> <p>Random verb: video of a man <b>airlifting</b> on his phone</p> <p>Antonym verb: <b>[no antonym]</b></p> <p>T5: video of a man <b>surfing the web</b> on his phone</p> <p>PaLM: video of a man <b>taking a selfie</b> on his phone</p>

Figure A.4. **PaLM vs. rule-based methods for hard negative generation:** We compare hard verb negative caption generation using PaLM to T5 and rule-based methods such as replacing detected verbs by random verbs or antonym verbs. We observe that randomly changing the verb often results in sentences which are linguistically and semantically incorrect, and that antonym verbs are often not present in NLTK [8]. On the other hand, LLM based methods such as T5 and PaLM result in meaningful sentences. We note that LLM based methods can change more than just the verb: in the last row, replacing ‘texting’ by ‘surfing the web’ with T5 and ‘taking a selfie’ with PaLM. In some cases, this can make it an easier negative: for example, in the third row, replacing ‘sitting’ by ‘sitting with a woman’ with T5.

cedure (Sec. C.5), the T5 hard negative generation process (Sec. C.6), and the Kinetics-verb split we propose (Sec. C.7).

### C.1. Baselines

We describe in more detail baselines presented in the main paper for MSR-VTT, NEXT-QA, Kinetics-400 and SVO-Probes.

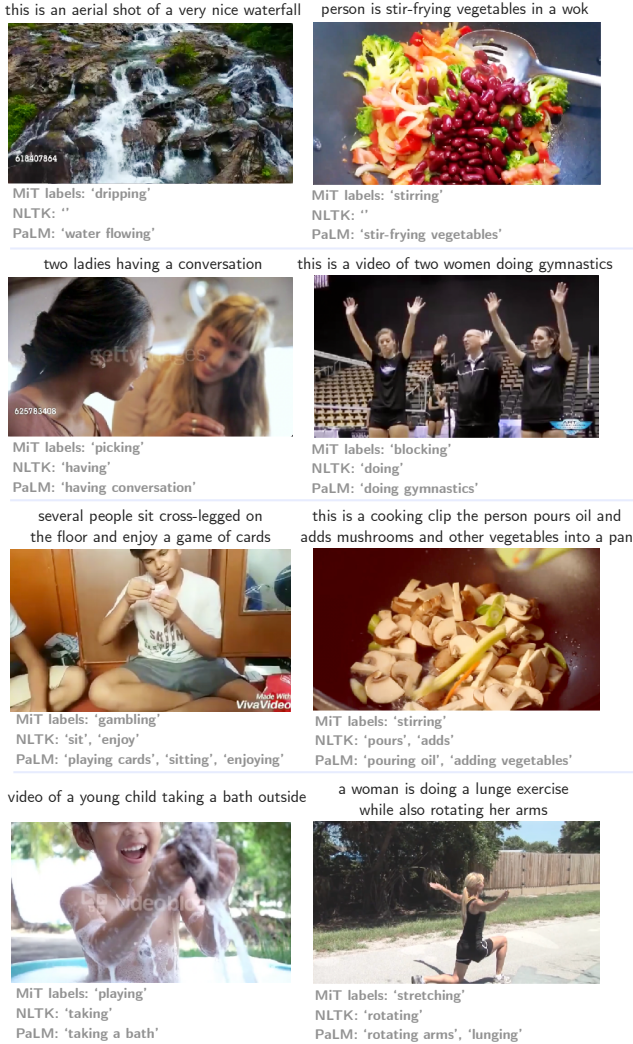


Figure A.5. **PaLM vs. rule-based methods for verb phrase extraction:** We compare verb phrase extraction using PaLM to: (i) using action labels for clips from the Moments in Time (MiT) dataset and (ii) using a rule-based method such as NLTK [8] to isolate verbs. In the top row, we show examples where NLTK outputs no label as a verb is not present in the sentence (first row, left) or is not detected (first row, right). In the second row, we show examples where extracting verbs with NLTK (e.g. ‘doing’, ‘having’) does not convey crucial information for understanding the action in the video. In the last two rows, we show examples where the MiT labels conceal valuable fine-grained action information in the video, whereas PaLM can recover this from the caption: (third row, left) the video is labelled as ‘gambling’, PaLM extracts ‘playing cards’; (third row, right) the video is labelled as ‘stirring’, PaLM extracts ‘pouring oil’ and ‘adding vegetables’; (last row, left) the video is labelled as ‘playing’, PaLM extracts ‘taking a bath’; (last row, right) the video is labelled as ‘stretching’, PaLM extracts ‘rotating arms’ and ‘lunging’. Overall, our PaLM method of extracting verbs from captions performs best qualitatively and quantitatively (as shown in Tab. 5 (right) of the main paper).

**MSR-VTT.** We show the performance of VideoCLIP [84], CLIP [58] and InternVideo [75] zero-shot. VideoCLIP trains a transformer for video and text by contrasting temporally overlapping positive video-text pairs with hard negatives from nearest neighbor retrieval. More details for the CLIP baseline can be found in [58]. InternVideo explores jointly using masked video modelling and video-language contrastive learning as pretraining objectives. In the fine-tuned setting, we compare to ClipBERT [38], MMT [26], VideoCLIP [84], C4CL-mP [53]. ClipBERT focuses on sparse training to reduce video processing overhead and applying image-text pretraining for video-text tasks. MMT uses a multi-modal transformer to encode video and BERT [17] for text. C4CL-mP corresponds to the CLIP4CLIP [44] reimplementation by Park et al. [53] with just mean pooling (without any Transformer Encoder for temporal modelling of frames).

**NEXT-QA.** We show the performance of CLIP zero-shot. More details for this baseline can be found in [58]. In the fine-tuned setting, we compare to HGA [33]: a deep heterogenous graph network which aligns inter- and intra-modality information (appearance, motion and text) to reason and answer the question. We also compare to ATP, Temp[ATP] and Temp[ATP]+ATP from [9]. ATP consists of a Transformer which learns to select a single (frozen) CLIP frame embedding from a video, given the sequence of video frame embeddings and question embedding, for the task of video question answering. For training, they use a cross entropy loss over the answer set. Temp[ATP] is an extension of ATP, where the video is first partitioned into  $k$  clips, and a single frame embedding is selected using ATP from each clip. These  $k$  frame embeddings are then aggregated to a video-level representation using a Transformer, before being passed to the downstream task. Temp[ATP]+ATP corresponds to an ensemble of both ATP and Temp[ATP]. Finally we compare to VGT [83], which consists of a video graph transformer that explicitly encodes objects, relations and dynamics. VGT also uses disentangled video and text Transformers to better measure relevance between video and text.

**Kinetics-400.** We show the performance of Flamingo [3], ActionCLIP [74] and CLIP [58] zero-shot. Flamingo is a visual-language model, which leverages pretrained vision and language models and bridges them effectively by using gated cross-attention and dense layers. ActionCLIP [74] reformulates action recognition into a video-text matching problem within a multimodal contrastive learning framework. More details for CLIP can be found in [58].

**SVO Probes.** We show the performance of No-MRM-MMT (the best performing model in [31]) and CLIP [58] zero-shot. No-MRM-MMT corresponds to a multi-modal transformer (similar to the ViLBERT [43] architecture) with a masked language modeling loss (MLM),



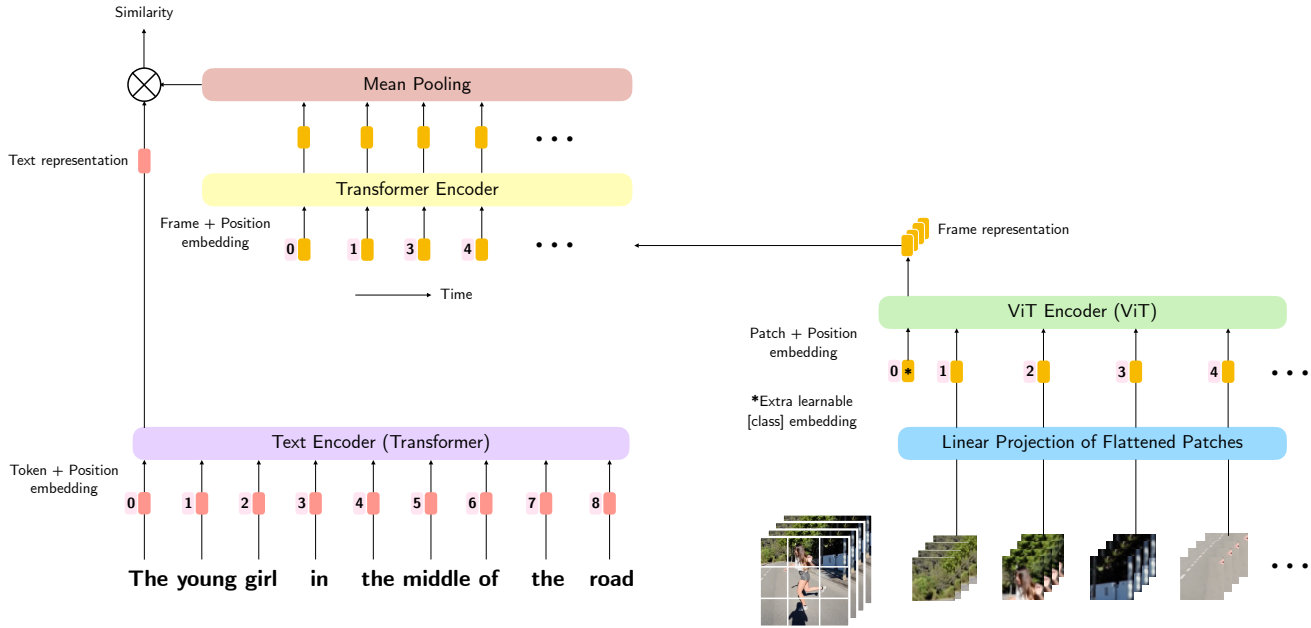


Figure A.6. **CLIP4CLIP Architecture:** Figure adapted from [44]. The model consists of a video encoder, text encoder and similarity calculator. Each frame is passed through ViT to obtain a frame representation at the output of the [class] token. The  $T$  frame representations are then passed through a Transformer for sequence modelling and averaged with a mean pooling operation to obtain a video-level representation. The video representation is then compared to the text representation through the cosine similarity.

an image-text matching loss (ITM) that classifies if an image-sentence pair are matching, but no masked region modeling loss (MRM). More details for the CLIP baseline can be found in [58].

## C.2. CLIP4CLIP Architecture

Here, we describe the CLIP4CLIP network architecture [44], illustrated in Figure A.6, used in all our experiments. This architecture consists of three components: a video encoder, a text encoder, and a similarity calculator. We describe each component in detail next. We note that all three components are fine-tuned in all our experiments.

**Video encoder.** The pretrained CLIP (ViT-B/32) [58] image encoder is used to obtain frame representations. Specifically, video frames are first sampled from the video and reshaped into a sequence of flattened 2D patches. These patches are then linearly projected to 1D tokens before being inputted to ViT [20], a 12-layer Transformer. The output from the [class] token is used as the video frame representation: given  $T$  input frames, we obtain  $T$  frame representations. In practice, we select 32 frames (with initial resolution  $256 \times 256$ , of which augmented crops of size  $224 \times 224$  are taken) in a video at 25fps, at a stride of 2 frames for training.

**Text encoder.** The CLIP pretrained text encoder is used to embed the caption. It corresponds to a 12-layer Transformer model; further details can be found in [58].

**Similarity calculator.** The goal is to learn a function to calculate the similarity between video-text pairs inputted to the model in such a way that video-text pairs which *match* have a high similarity, and otherwise have a low similarity. Therefore, we ultimately want to compare a text and video-clip representation. The ViT encoder outputs a representation for each of the sequence of frames without any temporal modelling. We therefore first pass these frame embeddings (along with temporal positional embeddings) through a 4-layer Transformer encoder. We then apply a mean-pooling operation to the new frame embeddings to obtain a video-level representation. Finally, we calculate the cosine similarity between the video and text representations.

Following the protocol in [44], the positional embeddings in the similarity calculator are initialised by repeating the position embedding from CLIP’s text encoder. The Transformer encoder is initialised by the corresponding layers’ weight of the pretrained CLIP image encoder. The rest is randomly initialised.

## C.3. Fine-tuning details

**MSR-VTT.** When fine-tuning on MSR-VTT, we use the 9K and 7K training split for the retrieval and multi-choice settings respectively. For the 9K split, we train for 100 epochs with a base learning rate of  $1e-7$ , a weight decay of  $1e-2$  and temperature of  $5e-3$ . For the 7K split, we train for 100

epochs with a base learning rate of  $1e-7$ , a weight decay of  $1e-2$  and temperature of  $5e-3$ . For both settings, we train with the hard negative contrastive loss and discard the verb phrase loss. Indeed, we use PaLM to generate hard negative captions for MSR-VTT, since it is a video-text retrieval dataset, similarly to SMiT. We sample 32 frames per video at 25 fps with a stride of 14.

**NEXT-QA.** For fine-tuning on NEXT-QA, we concatenate the question and answer pairs before passing them through the CLIP4CLIP text tower. We continue using the hard-negative cross-modal contrastive loss during fine-tuning, treating the four incorrect question-answer pairs as hard negatives. We discard the verb phrase loss. We train for 100 epochs with a base learning rate of  $1e-6$ , a weight decay of  $5e-2$  and temperature of  $1e-3$ . We maintain a batch size of 256. We sample 32 frames per video at 25 fps with a stride of 24.

#### C.4. Evaluation protocols

**MSR-VTT.** For the standard setting, we evaluate text-to-video retrieval (R@1) on the 1K validation split and 3K Random MC. In the former, the model must associate the text to the correct video, among 1000 videos. For the latter, the model must associate the video to the right caption, among 5 captions, where the 4 negative captions are randomly chosen from other videos. For our verb-focused setting, we use the Verb<sub>H</sub> multiple choice (MC) validation split from [53]. Verb<sub>H</sub> MC covers a subset of the videos in the 3K Random MC split, with 2,554 video-text instances, but the task is harder. In the Verb<sub>H</sub> MC setting, one of random negative captions is replaced by a *hard verb negative*, where the correct sentence’s verb has been modified manually in such a way that the new sentence is inconsistent with the video. We mark the model prediction as correct if the ground truth sentence among the 5 captions has the highest similarity score with the video. We sample 32 frames per video at 25 fps with a stride of 14.

**Kinetics-400.** We follow [58] to evaluate classification in a zero-shot setting: we feed in all class labels (without any prompt) to the text tower and mark the prediction as correct if the correct label has the highest similarity with the video. For the ‘Kinetics-verb’ split, we restrict the evaluation to 97 classes which we manually identify as requiring verb understanding (see Sec. C.7). We note that we still feed in all 400 class labels for measuring classification on Kinetics-verb. We sample 32 frames per video at 25 fps with a stride of 14.

**NEXT-QA.** We concatenate the question and answer pairs before passing them through the CLIP4CLIP text tower. We mark the model prediction as correct if the correct question-answer pair among the 5 options has the highest similarity score with the video. We sample 32 frames per video at 25 fps with a stride of 24.

**SVO probes.** This is an image-text benchmark [31], specif-

ically designed to measure progress in verb understanding. We evaluate our baseline and VFC framework on a subset of 12,936 images from the original 14,102 images since some images are no longer accessible (the corresponding urls are corrupted). In [31], the authors calculate the accuracy of positive and negative image-text pairs: they pass image-text pairs through their model and label an image-sentence pair as negative if the classifier output is  $< 0.5$  and positive otherwise. Our model confidences are calibrated differently, therefore we instead report Average Precision (AP). To evaluate on this dataset, we simply replicate the image 32 times as input to our video model.

#### C.5. PaLM prompting

**PaLM hard negative generation.** We include below our full prompt template for automatic generation of hard negatives. We insert the caption for which we want to generate hard verb negatives at `{input caption}`.

---

*In this task, you are given an input sentence. Your job is to tell me 10 output sentences with a different meaning by only changing the action verbs.*

*Input: A man walks up to a woman holding an umbrella in a garden.*

*Outputs:*

- 1) A man jumps up to a woman throwing an umbrella in a garden.*
- 2) A man runs up to a woman opening an umbrella in a garden.*
- 3) A man walks away from a woman buying an umbrella in a garden.*
- 4) A man throws up on a woman carrying an umbrella in a garden.*
- 5) A man punches a woman swinging an umbrella in a garden.*
- 6) A man sits with a woman wrapping up her umbrella in a garden.*
- 7) A man talks to a woman closing an umbrella in a garden.*
- 8) A man flirts with a woman playing with an umbrella in a garden.*
- 9) A man skips to a woman leaning on her umbrella in a garden.*
- 10) A man sprints to a man losing her umbrella in a garden.*

*Input: Surfers ride the waves in an ocean. Outputs:*

- 1) Surfers get hit by the waves in an ocean.*
- 2) Surfers swimming in the waves in an ocean.*
- 3) Surfers meditating by the waves in an ocean.*
- 4) Surfers drowning in the waves in an ocean.*
- 5) Surfers asking for help in the waves in an ocean.*
- 6) Surfers teaming up in the waves in an ocean.*
- 7) Surfers snorkeling in the waves in the ocean.*
- 8) Surfers taking photos by the waves in the ocean.*
- 9) Surfers getting ready to go into the waves in the ocean.*

10) *Surfers stretching by the waves in the ocean.*

*Input: A dentist holds the replica of a human mouth he shows how important flossing your teeth is.*

*Outputs:*

1) *A dentist cleans the replica of a human mouth he presents how unimportant flossing your teeth is.*

2) *A dentist breaks the replica of a human mouth he screams how important flossing your teeth is.*

3) *A dentist fixes the replica of a human mouth he says how important flossing your teeth is.*

4) *A dentist buys the replica of a human mouth he explains how important brushing your teeth is.*

5) *A dentist plays with the replica of a human mouth he remembers about how important washing your teeth is.*

6) *A dentist tidies the replica of a human mouth he rambles on about how important breaking your teeth is.*

7) *A dentist rotates the replica of a human mouth he presents how important fracturing your teeth is.*

8) *A dentist places on his legs the replica of a human mouth he shows how important flossing your teeth is.*

9) *A dentist searches for the replica of a human mouth he shows how important grinding your teeth is.*

10) *A dentist picks up the replica of a human mouth he presents how important whitening your teeth is.*

*Input: Looks like a band playing on the stage and perhaps Community Center and people gathered around watching.*

*Outputs:*

1) *Looks like a band fighting on the stage and perhaps Community Center and people gathered around crying.*

2) *Looks like a band dancing on the stage and perhaps Community Center and people gathered around smiling.*

3) *Looks like a band singing on the stage and perhaps Community Center and people gathered around filming.*

4) *Looks like a band bowing on the stage and perhaps Community Center and people gathered around clapping.*

5) *Looks like a band making a speech on the stage and perhaps Community Center and people gathered around listening.*

6) *Looks like a band laughing on the stage and perhaps Community Center and people gathered around cheering.*

7) *Looks like a band working on the stage and perhaps Community Center and people gathered around standing.*

8) *Looks like a band holding hands on the stage and perhaps Community Center and people gathered around praying.*

9) *Looks like a band jumping on the stage and perhaps Community Center and people gathered around encouraging.*

10) *Looks like a band yelling on the stage and perhaps Community Center and people gathered around watching.*

*Input: {input caption}*

*Outputs:*

---

**PaLM verb phrase extraction.** We use PaLM to extract verb phrases from the original caption, where a verb phrase can correspond to a single verb or a verb-noun pair depending on the caption. We use PaLM-540B with output sequence length 256, beam size of 4, and temperature of 0.2. We post-process the outputs by removing text after any newline character. We include our full prompt template for automatic extraction of verb phrases below. We insert the caption for which we want to extract a verb phrase at **{input caption}**.

---

*In this task, you are given an input sentence. Your job is to output the action verb phrases.*

*Input: the young girl in the middle of the road she is dancing.*

*Output: ['dancing']*

*Input: a city area can be seen that has people in the walkways of runways.*

*Output: []*

*Input: this is a video of a birthday and she has a green colored dress and they are cutting a cake there's a clown on the side and the parents seem to be clap.*

*Output: ['cutting cake', 'clapping']*

*Input: one woman is talking to the camera about being safe he has a shirt with pal pal on it in the greenery behind her.*

*Output: ['talking to camera']*

*Input: a bicycle with a specialized back wheel slides along a wet paper.*

*Output: ['sliding']*

*Input: a person clicking an object that is connected to a speaker.*

*Output: ['clicking']*

*Input: it's a video of a football game and one of the blue team is throwing the football really far into the endzone.*

*Output: ['throwing football']*

*Input: this is a video of someone filing their nails.*

*Output: ['filing nails']*

*Input: airplane with the words British Airways can be seen over top.*

*Output: []*

*Input: man sitting standing at the front of the room is giving speech and asking an audience if they've ever heard of a specific song.*

*Output: ['standing', 'giving speech', 'asking']*

*Input: it shows a video of a man talking on the phone yeah glasses and has a black phone.*

*Output: ['talking on phone']*

*Input: hitchhiker is on the side of the road by a truck stop pulling a sign that says North.*

*Output: ['pulling a sign']*

*Input: this is a video of a man on a ladder the man is cutting down a tree branch the man is wearing red.*

*Output: ['cutting tree']*

*Input: on an indoor gym on a hard Brown meth there's a man young man with a barbell with lots of heavy weights on each side and he has it over his head stiff arm straight arm going to be and then he drops it on the floor while he does so you can hear the clanking of the weight that they smack against each other.*

*Output: ['dropping']*

*Input: he is using a large chainsaw to cut inside of a tree branch.*

*Output: ['cutting tree']*

*Input: I meant stacking up his cups for cup stacking concentration for a party.*

*Output: ['stacking cups']*

*Input: a large field shown with garbage and water flowing through it.*

*Output: ['water flowing']*

*Input: a washing machine washes the clothes.*

*Output: ['washing clothes']*

*Input: {input caption}*

*Output:*

---

**PaLM positive generation.** We use PaLM to generate positive sentences where the verb in the original caption is changed to a synonym verb, but the remaining context is unchanged. We use PaLM-540B with output sequence length 512, beam size of 1, and temperature of 0.7. We post-process the outputs by removing text after any newline character and by filtering out candidates which contain the same verbs as the original caption. We include our full prompt template for automatic generation of positives below. We insert the caption for which we want to generate a positive sentence at **{input caption}**.

---

*In this task, you are given an input sentence. Your job is to tell me 10 output sentences with the same meaning by only changing the action verbs.*

*Input: A man walks up to a woman holding an umbrella in a garden.*

*Outputs:*

*1) A man strolls up to a woman holding an umbrella in a garden.*

*2) A man marches up to a woman holding an umbrella in a garden.*

*3) A man strides up to a woman holding an umbrella in a garden.*

*4) A man wanders up to on a woman carrying an umbrella in a garden.*

*5) A man tramps up to a woman holding an umbrella in a garden.*

*6) A man steps up to with a woman holding an umbrella in a garden.*

*7) A man wanders up to a woman holding an umbrella in a garden.*

*8) A man treads up to a woman holding an umbrella in a garden.*

*9) A man truges up to a woman holding an umbrella in a garden.*

*10) A man treaks to a woman holding her umbrella in a garden.*

*Input: A dentist holds the replica of a human mouth he shows how important flossing your teeth is.*

*Outputs:*

*1) A dentist grasps the replica of a human mouth he shows how important flossing your teeth is.*

*2) A dentist carries the replica of a human mouth he shows how important flossing your teeth is.*

*3) A dentist clutches the replica of a human mouth he shows how important flossing your teeth is.*

*4) A dentist grips the replica of a human mouth he shows how important flossing your teeth is.*

*5) A dentist holds the replica of a human mouth he explains how important flossing your teeth is.*

*6) A dentist holds the replica of a human mouth he presents how important flossing your teeth is.*

*7) A dentist holds the replica of a human mouth he demonstrates how important flossing your teeth is.*

*8) A dentist holds the replica of a human mouth he communicates how important flossing your teeth is.*

*9) A dentist holds the replica of a human mouth he displays how important flossing your teeth is.*

*10) A dentist holds the replica of a human mouth he highlights how important flossing your teeth is.*

*Input: This is a video of somebody touching wood.*

*Outputs:*

*1) This is a video of somebody tapping wood.*

*2) This is a video of somebody stroking wood.*

*3) This is a video of somebody pressing wood.*

*4) This is a video of somebody handling wood.*

*5) This is a video of somebody patting wood.*

*6) This is a video of somebody brushing wood.*

*7) This is a video of somebody grazing wood.*

*8) This is a video of somebody poking wood.*

*9) This is a video of somebody caressing wood.*

*10) This is a video of somebody gripping wood.*

*Input: This is a video of a group of adults outside dancing.*

*Outputs:*

*1) This is a video of a group of adults outside whirling.*

*2) This is a video of a group of adults outside twirling.*

*3) This is a video of a group of adults outside swaying.*

*4) This is a video of a group of adults outside partying.*

*5) This is a video of a group of adults outside getting down.*

*6) This is a video of a group of adults outside spinning.*

*7) This is a video of a group of adults outside bouncing.*

*8) This is a video of a group of adults outside bopping.*

*9) This is a video of a group of adults outside waltzing.*

*10) This is a video of a group of adults outside prancing.*

Input: {input caption}

Outputs:

---

## C.6. T5 generations

As well as using PaLM to generate hard verb negative captions, we experiment with using a bidirectional language model, T5-Base [59]: a 220 million parameter encoder-decoder Transformer. It is pretrained on the Colossal Clean Crawled Corpus (C4) [23] on a multi-task mixture of unsupervised and supervised tasks, with all tasks being converted into a text-to-text format. T5 is trained with a Masked Language Modelling (MLM) loss, similarly to BERT [17], with minor differences. MLM involves masking certain tokens in an input sequence before passing them to the model, and tasking the model with predicting the masked spans.

As T5 has been trained with a span-mask denoising objective, we use it at inference time in *cloze* form (fill in the blanks) to replace words in captions by targeted masking. Specifically, our method consists of the following steps:

(1) **Verb Identification:** we start by identifying verbs in text captions, leveraging PoS tagging with NLTK [8].

(2) **T5 prediction:** We then replace the verb tokens with a [MASK] token, and feed the masked sentence to T5. We keep the Top- $K$  phrases predicted by the model (with  $K = 50$ ). Unlike [53], we do not fine-tune T5 for verb modelling specifically, but rather use it in a zero-shot setting, which we find is sufficient to generate plausible negatives.

(3) **Negatives Filtering:** The  $K$  candidate sentences are then filtered to remove sentences which contain the same verbs as the original caption.

## C.7. Kinetics-verb

In order to assess our method’s true verb understanding in the downstream task of action classification, we introduce ‘Kinetics-verb’: a subset of 97 classes from Kinetics-400 [11] where we isolate classes that share a *common noun* with another class, but have a *different verb* (and therefore action). We include the set of 97 classes below:

[**hair**: braiding hair, brushing hair, curling hair, dying hair, fixing hair, washing hair, getting a hair cut; **nails**: doing nails, cutting nails; **legs**: waxing legs, massaging legs, shaving legs, stretching leg, swinging legs; **hands**: washing hands, shaking hands, **arm**: stretching arm, exercising arm, arm wrestling; **watermelon**: cutting watermelon, eating watermelon; **floor**: mopping floor, cleaning floor, sanding floor, sweeping floor; **baby**: baby waking up, carrying baby, crawling baby; **back**: waxing back, bending back, massaging back; **feet**: massaging feet, washing feet; **dog**: walking the dog, grooming dog, training dog; **cake**: eating cake, making a cake; **guitar**: strumming guitar, playing

guitar, tapping guitar; **cards**: shuffling cards, playing cards; **present**: wrapping present, opening present; **egg**: cooking egg, egg hunting, scrambling eggs; **shoes**: shining shoes, cleaning shoes; **pool**: cleaning pool, jumping into pool; **snow**: biking through snow, shoveling snow; **rope**: skipping rope, climbing a rope; **fish**: catching fish, feeding fish; **eyebrows**: filling eyebrows, waxing eyebrows; **computer**: using computer, assembling computer; **tree**: climbing tree, planting trees, trimming trees; **car**: driving car, pushing car; **golf**: golf chipping, golf driving, golf putting; **beer**: drinking beer, tasting beer; **horse**: grooming horse, riding or walking with horse; **paper**: folding paper, ripping paper, shredding paper; **fire**: extinguishing fire, juggling fire; **head**: shaking head, shaving head; **water**: surfing water, water skiing, water sliding; **ice**: ice climbing, ice fishing, ice skating; **basketball**: dunking basketball, dribbling basketball, playing basketball, shooting basketball; **finger**: drumming fingers, finger snapping; **baseball**: catching or throwing baseball, hitting baseball; **soccer ball**: juggling soccer ball, kicking soccer ball]