LocoMotion: learning motion-focused video-language representations

Doughty, H.R.; Thoker, M.F.; Snoek, M.G.C.; Cho, M.; Laptev, I.; Tran, D.; ... ; Zha, H.

LocoMotion: Learning Motion-Focused Video-Language Representations

Hazel Doughty[1(✉)], Fida Mohammad Thoker[2,3], and Cees G. M. Snoek[2]

[1] Leiden University, Leiden, The Netherlands
[2] University of Amsterdam, Amsterdam, The Netherlands
[3] King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Abstract. This paper strives for motion-focused video-language representations. Existing methods to learn video-language representations use spatial-focused data, where identifying the objects and scene is often enough to distinguish the relevant caption. We instead propose LocoMotion to learn from motion-focused captions that describe the movement and temporal progression of local object motions. We achieve this by adding synthetic motions to videos and using the parameters of these motions to generate corresponding captions. Furthermore, we propose verb-variation paraphrasing to increase the caption variety and learn the link between primitive motions and high-level verbs. With this, we are able to learn a motion-focused video-language representation. Experiments demonstrate our approach is effective for a variety of downstream tasks, particularly when limited data is available for fine-tuning. Code is available: https://hazeldoughty.github.io/Papers/LocoMotion/.

Keywords: Video-Language Representation · Video Understanding

1 Introduction

Recent video-language models [8,10,38,48,80,84] have shown impressive performance on a wide range of video-language tasks. The increased success in recent years is heavily linked to the creation of large-scale web-scraped video-language datasets [4,51,84] and the ability to bootstrap training with more readily available image-text pairs [4]. Due to this collection process, captions in video-language datasets are focused on the spatial aspects of the video, *i.e.* the scene and objects. For instance, in Fig. 1 understanding concepts like *group of people*, *kites*, *beach*, and *playing* are key to determining this is the caption corresponding to the video. All of these concepts can be understood from any frame of the video meaning that many popular video-language benchmarks [1,32,35,39,79,81] can even be solved by only looking at a single frame [6,37]. Hence, models trained

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-96-0908-6_1.

M. Cho et al. (Eds.): ACCV 2024, LNCS 15474, pp. 3–24, 2025.
https://doi.org/10.1007/978-981-96-0908-6_1

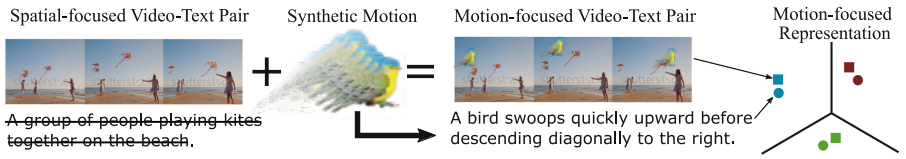

Fig. 1. Motion-Focused Video-Language Representations. Video-text pairs from web data have a spatial focus. To learn motion-focused video-language representations, we create motion-focused video-text pairs for pre-training by adding synthetic motions to videos and automatically generating new captions describing these motions.

on existing video-language data struggle to understand motion, making their learned representations unsuitable for motion-focused downstream tasks.

In this paper, we remove the spatial focus of video-language representations and propose LocoMotion which instead trains representations to have a motion focus. We achieve this by discarding the spatial-focused captions and generating new motion-focused captions for pre-training. However, we cannot generate motion-focused captions without knowing the motions present in the video. We thus generate synthetic local object motions that we inject into the pre-training videos. This also increases the variety of motions the videos contain. Since we know the properties of these local object motions, we can automatically generate captions describing them. Figure 1 shows an illustration of our approach.

We make three contributions. First, we highlight the spatial focus of captions in current video-language pre-training datasets and shift the emphasis to learning representations for motion-oriented tasks. Second, we propose a method to learn motion-focused video-language representations from synthetic local object motions alongside automatic descriptions of these motions. Third, we propose verb-variation paraphrasing to provide a much more varied and more high-level way to describe the motion primitives present in our local object motions. Experiments demonstrate the advantage of our approach to different motion-focused downstream tasks, particularly when limited fine-tuning data is available.

2 Related Work

Video-Language Representation Learning. aims to learn a joint representation for videos and text which can easily be adapted to downstream tasks. Methods are typically trained with multiple video frames [4,10,38,42,48,76,80,84,90,92]. For instance, Merlot [84] learns to match captions to frames and predicts the correct frame order. CLIP4CLIP [48] uses 3D space-time patches when transferring CLIP [61] to video. VindLU [10] introduces a recipe for effective pre-training. Despite progress in the ability of video-language models to represent temporal data, recent works [6,37] show single-frames are sufficient for many video-language datasets [1,32,35,39,79,81]. Singularity [38] achieves impressive performance by training on randomly sampled single frames while ensembling

multiple frames at inference. ATP [6] proposes an a temporal probe to select the best frame for inference. These works highlight the spatial focus of current video-language models and benchmarks. We rectify this spatial focus, by learning motion-focused representations suitable for video-language tasks.

Two other works also aim to enhance the temporal understanding of video-language models [2,77]. Bagad *et al.* [2] instill video language models with the concept of before/after relations. Wang *et al.* [77] patch in action knowledge to video-language models in the form of sensitivity to video reversal and distinguishing actions with their antonyms. Differently to these works, we aim to enhance video-language representations with an understanding of motion.

3D Human Motion Retrieval. Several works learn motion-focused representations from human pose sequences and text [9,23,25,57,58,69,87]. Most generate 3D human motion from text, although retrieval is used for evaluation [25]. Petrovich *et al.* [58] were first to tackle text-to-3D human motion retrieval as a standalone task, extending text-to-motion synthesis [57] with a contrastive loss to structure the cross-modal latent space. Liu *et al.* [47] instead bridge the gap between action semantics and motion representations with kinematic hierarchies to indicate the evolution of joint positions and limb orientations through actions. Rather than learn from 3D human motion, we aim to learn a motion-focused video-text representation from the (partly synthesized) pixel space.

Supervised Fine-Grained Motion Learning. While few works study motion-focused video-language representations, much progress has been made in motion learning in a unimodal setting. Approaches distinguish fine-grained actions with specialized network blocks [34,36,45,49], separating motion and appearance [44,67], aggregating temporal scales [20,53,83], and obtaining mid-level representations with sparse coding [50,59,64]. Other works use input video representations that better highlight motion such as skeleton data [18,28] and optical flow [21,66]. Several works go a step beyond distinguishing fine-grained actions and instead identify motion differences between instances of the same action with adverbs [16,17,52], action attributes [85] and repetitions [29,86,88]. Different from all these works, we learn a motion-sensitive video-language representation.

Self-Supervised Motion Representation Learning. While self-supervised learning in video typically targets coarse-grained actions [12,30,33,46,55,56,60,68], many recent works explore fine-grained actions that require temporal and motion understanding [15,22,27,31,54,70,71,73,74,78]. Some works [15,74] achieve motion-focus by removing background bias [15,74], using optical flow [22,27,54,78] or masked autoencoders [19,72]. Most relevant to us are works that inject motion into videos [31,71,73]. These works learn representations by predicting properties and trajectories of injected motions [31,73] or by contrasting videos with different motion [71]. We take inspiration from these works and generate motion-focused video captions corresponding to injected local object motions. This allows us to learn a motion-focused video-language representation without manual motion descriptions.

3 Spatial Focus of Video-Language Representations

We first explore the suitability of current video-language datasets for learning motion concepts by examining their collection processes and caption contents.

Definition and Notation. Video-language models aim to learn representations that indicate how well individual captions in a set of texts T describe the content of videos in the set V. Specifically, given a video $v \in V$ with corresponding text description $t \in T$, the goal is to learn a function $\mathcal{F}(v, t)$ that outputs the similarity of video v to caption t. The similarity of video v should be higher with ground-truth caption t than with any other caption t'. Likewise, $\mathcal{F}(v, t)$ should be higher than $\mathcal{F}(v', t)$ for videos v' unrelated to t. This representation can then be applied to new data at inference or fine-tuned for a specific domain or task. The success of this learned representation therefore heavily relies on the contents of the videos V and captions T. If concepts are not present in T then the model has little hope of learning their visual representation regardless of the model capabilities. Moreover, concepts present in T may not be needed to match video-text pairs, again resulting in the representation not learning these concepts. It is possible to learn additional concepts during fine-tuning, however this requires large amounts of annotated data for the downstream task.

Video-Language Datasets. We examine the contents of the popular pre-training datasets WebVid [4], HowTo100M [51], YT-Temporal [84], InternVid [76], and CMD [3] and downstream datasets: MSR-VTT [81], DiDeMo [1], VATEX [75], Charades [65] and ActivityNet Captions [7]. Except Charades, where people record scripted actions, all these datasets collect videos from sites such as Flickr and YouTube. Captions are collected by scraping the video's alt-text [4] or description [3], automatic speech recognition [51,84], image captioning [76] and manual annotation [1,7,75,81]. In all caption creation methods, the text for each video is created in isolation, without accounting for what makes this video different from others in the dataset. This results in the captions ignoring fine-grained differences such as motion and instead focusing on the scene and objects.

Caption Content. We verify the spatial focus of these dataset's captions by exploring the parts-of-speech used. Specifically in Fig. 2 we display the average number of nouns, adjectives, verbs, adverbs, and adpositions per caption. For all datasets, nouns are the largest component of each caption, demonstrating the spatial focus of these video-language datasets. This is particularly noticeable for pre-training dataset WebVid where the second and third most common word types are adpositions and adjectives which also do not require motion or temporal understanding. Verbs and adverbs which require motion understanding occur much less frequently. Furthermore, large proportions of the captions in these datasets are uniquely identifiable from the nouns. We identify this by finding the set of nouns in a caption and counting unique sets. For instance, although YT-Temporal and CMD have a better ratio of verbs to nouns than other datasets, a large proportion of captions (87% and 82% respectively) can be matched to the correct video clip by only recognizing the correct nouns. The exception to this

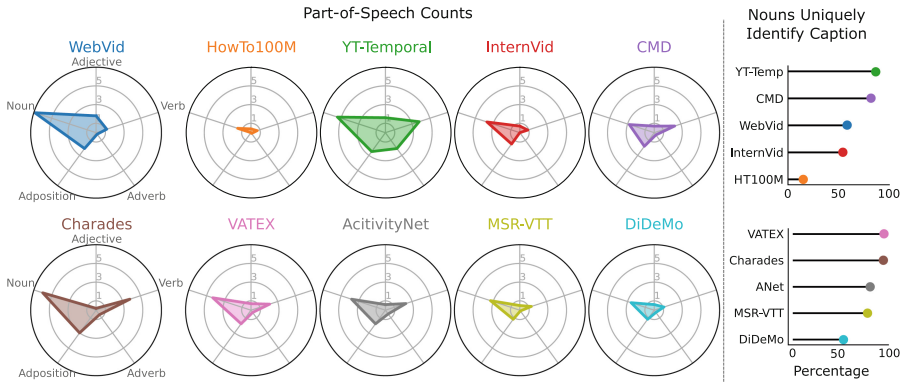

Fig. 2. Caption Content of video-language datasets. Left: We show the average number of nouns, adjectives, verbs, adverbs, and adpositions per captions for 5 popular pre-training (top) and downstream datasets (bottom). Right: The percentage of captions uniquely identifiable using only the nouns. Current video-language datasets have a spatial focus demonstrated by the average number of nouns each caption contains and the percentage of captions that can be uniquely identified by only their nouns.

is HowTo100M, where automatically generated subtitles are divided into short phrases. This presents a different issue as a large proportion of the words in the captions do not have any visual correspondence in the video.

Captioning with VLMs. A simple way to create new captions for existing video-language datasets is with a VLM, *e.g.* [41]. However, training with such video-text pairs would not result in a motion-focused representation. First, captions are limited by the amount of motion in the original videos which is often small. Second, due to the spatial-focus of video-text pretraining datasets, VLMs are also spatial-focused. This is shown in Fig. 3 where we ask VideoChat2 [41] to describe the motion in this video for WebVid data. We tried various prompts and found motion descriptions are often absent, *e.g.* the orchids swaying left and right are missed, or are incorrect *e.g.* 'strolling'.

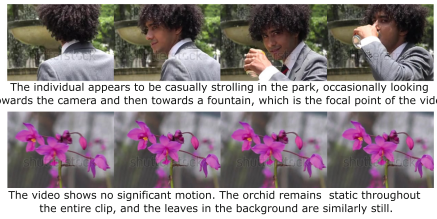

Fig. 3. Captioning with VLMs also results in a spatial-focus. Examples generated with VideoChat2.

Conclusion. We conclude current video-language pre-training has a strong spatial focus due to the datasets used. This is not revealed by downstream datasets, as they also have a spatial focus. The cause of these issues are the captions and the aspects of the videos these captions describe. This cannot be simply solved by using VLMs to generate new captions as these are also spatial-focused. Thus, we propose an automated method to create motion-focused video-text pairs from existing datasets and use this to pre-train a video-language representation.

4 LocoMotion: Learning with Local Object Motion

We learn motion-focused video-language representations with self-supervision. Towards this, we propose LocoMotion (Fig. 4) which learns from local object motion and corresponding text descriptions. We first add local object motions to input videos (Sect. 4.1). We then generate new captions to describe these motions (Sect. 4.2), creating variety in our description through verb-variation paraphrasing (Sect. 4.3). Finally, we train a motion-focused video-language representation with our new video-text pairs (Sect. 4.4).

4.1 Motion Generation

As demonstrated in Sect. 3, prior works [6,37,38,42,48,80,84,92] learn from video-text pairs (v, t) focused on spatial aspects of video rather than motion. As a first step to learning motion-focused representations, we expand the motion variety of video v to v_{motion}. Specifically, we add an object to v and give this object motion by translating and rotating throughout the video. Although these motions may be unrealistic, adding such motions to input videos encourages the model to be more sensitive to motions and their temporal progression.

Appearance. To add a new motion to a video v with N frames $v{=}\{v_1, v_2, ..., v_N\}$, we first randomly select an object $o \in O$, where O is a set of segmented objects. We then sample a scale for this object such that the new height and width of the object $H'{\times}W'$ is less than the video's height and width $H{\times}W$. This object is then overlaid on the first video frame by uniformly sampling a starting position (x_1, y_1) for the center of the object such that $(\frac{H'}{2}, \frac{W'}{2}) \leq (x_1, y_1) \leq (H{-}\frac{H'}{2}, W{-}\frac{W'}{2})$ so that the object is contained in the video frame. This gives us an initial starting point from which we can give object o motion.

Translation. The first way we give object o motion is by translating it to different locations in the N video frames. We select K keyframes including the first and last frames (v_1 and v_N) as well as $K{-}2$ randomly selected frames. For each keyframe v_{k_i} we select a new location for the object center where $x_{k_{i-1}}{-}\delta(k_i{-}k_{i-1}) \leq x_{k_i} \leq x_{k_{i-1}}{+}\delta(k_i{-}k_{i-1})$ and similarly $y_{k_{i-1}}{-}\delta(k_i{-}k_{i-1}) \leq y_{k_i} \leq y_{k_{i-1}}{+}\delta(k_i{-}k_{i-1})$. This constrains the difference in object location between frames to be $< \delta$, thus giving the object o a smooth motion. Locations for non-keyframes are linearly interpolated between neighboring keyframes.

Rotation. We create a greater variety of more complex motions by including rotation. For each keyframe v_k we randomly select an angle θ_k. Similar to translation, angles are linearly interpolated between keyframes. We then apply rotation matrix $\left(\begin{smallmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{smallmatrix}\right)$ to object o before adding it to video frame v_i.

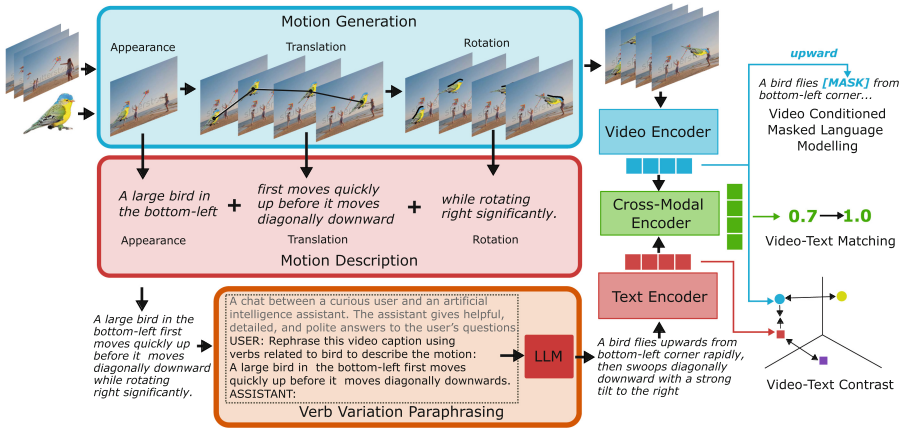

Fig. 4. LocoMotion learns a motion-focused video-language representation. Given a video, we randomly sample an object, a size, and a starting location giving the appearance of the initial frame. We generate local object motion by translating and rotating the object across time. As we know the motion parameters we can create corresponding text to accurately describe the object's motion in the video. Our verb-variation paraphrasing allows us to diversify the vocabulary and structure of this motion description and link low-level motion to more high-level verbs. Using the resulting video-text pairs to pre-train video, text and cross-modal encoders with masking, matching, and contrastive losses results in a motion-focused video-language representation.

4.2 Motion Description

Injecting motions into our training videos, not only increases the variety of motion but also means we know the exact motion of object o in video v_{motion}. With this knowledge, we can create a new video caption t_{motion} to describe the local motion of object o. Replacing the spatial-focused video caption t with t_{motion} encourages the pre-training to focus on correspondence between motion and text. We describe each part of the generated motion as follows, first using $K{=}2$ keyframes to generate object motion o before extending to $K{>}2$ keyframes.

Appearance. It is important to describe the appearance of object o so that the model can accurately locate the object to determine its motion. We consider three factors of object appearance: form, size and starting location. To describe each of these factors, we create several possible phrases that are applied depending on the parameters of object o. Appearance is described using the object's name in the set of objects O. Size is dependent on the sampled height and width of the object $H'{\times}W'$. If $H'W'{<}\alpha$ we describe the object as *small* and if $H'W'{>}\beta$ we describe it as *large*. For the initial position, we divide the frame into a three-by-three grid and describe the position of the object's center point (x, y) relative to this grid, *e.g. bottom-right*, *left*, *center*, *top-left*. We combine these descriptors as follows to give us a phrase $t_{appearence}$ focusing on the object's appearance:

$$t_{appearence} = \text{A} + \underbrace{\{\text{big}, \text{small}\}}_{\text{size of object } o} + \underbrace{\{\text{airplane}, \text{apple}, \ldots, \text{zebra}\}}_{\text{name of object } o} + \text{in the} + \underbrace{\{\text{top-left}, \text{top}, \ldots, \text{bottom-right}\}}_{\text{position of object } o} \quad (1)$$

We obtain phrases like *A large car in the left* or *A small piano in the top-right*.

Translation. Similar to the object's appearance, we can describe the translation of the object by combining predefined phrases. We describe three components of the translation motion: speed, distance, and direction. For speed we use *quickly*, *slowly*, or no qualifier depending on the average distance in the object's center between consecutive frames. We describe the distance moved as *a lot*, *a little*, or with no qualifier depending on the total distance moved between keyframes. To describe the direction we calculate the angle $\theta_{translate}$ between keyframes. This is quantized into four equal intervals with text descriptions: *upwards*, *left*, *downwards* and *right*. Additionally, we apply the modifier *diagonally* if $30 < |\theta_{translate}| \pmod{90} < 60$. With this, we can assemble the translation motion phrase as:

$$t_{translate} = \text{moves} + \underbrace{\{\text{quickly}, \text{slowly}\}}_{\text{speed}} + \underbrace{\{\text{diagonally}\}}_{} + \underbrace{\{\text{upwards}, \text{right}, \text{downwards}, \text{left}\}}_{\text{direction}} + \underbrace{\{\text{a lot}, \text{a little}\}}_{\text{distance}} \quad (2)$$

resulting in phrases such as *moves slowly diagonally left* or *moves upwards a lot*.

Rotation. We describe two aspects of rotation: direction and amount. For rotation direction, we use *left* or *right* depending on whether the rotation angle is positive or negative. To describe the rotation amount we use the rotation angle θ_k. If $\theta_k < \gamma$ we use *slightly*, if $\theta_k > \zeta$ we use *significantly*, otherwise we don't add a descriptor. The rotation phrase is:

$$t_{rotate} = \text{while rotating} + \underbrace{\{\text{left}, \text{right}\}}_{\text{rotation direction}} + \underbrace{\{\text{slightly}, \text{significantly}\}}_{\text{rotation amount}} \quad (3)$$

We can then assemble our motion-focused caption t_{motion} from the phrases:

$$t_{motion} = t_{appearance} + t_{translate} + t_{rotate}. \quad (4)$$

Multiple Keyframes. As we increase the complexity of the generated motion with $K > 2$ keyframes, we can also increase the complexity of the motion description to accurately describe each stage. To do this we first create an appearance

phrase $t_{appearance}$. We then create individual translation and rotation captions $t_{translate_i}$ and t_{rotate_i} to describe the motions between each pair of consecutive keyframes k_i and k_{i+1} as previously outlined. From this, we form an overall caption:

$$\begin{aligned} t_{motion} = t_{appearence} + \text{first} + t_{translate_1} + t_{rotate_1}, \\ + ... + \text{before it} + t_{translate_i} + t_{rotate_i}. \end{aligned} \tag{5}$$

Our motion description creates a variety of motion-relevant video captions; for each object, there are over 100 million possible motion descriptions. By using our motion-focused video-text pairs (v_{motion}, t_{motion}) instead of spatial-focused video-text pairs (v, t) a model can learn key motion concepts during pre-training.

4.3 Verb-Variation Paraphrasing

While our motion description creates a large number of motion-focused captions, the captions are formulaic. This presents two problems. First, the limited options for each motion concept make the masked language modeling loss less useful. Second, the difference in caption style between pretraining and downstream tasks creates a domain gap making it harder to adapt the model to the target task.

Paraphrasing with LLMs. We address these issues with LLM-based paraphrasing to gain variety in structure and vocabulary. We find the choice of LLM to be key. While most LLMs create diverse captions, they often remove or alter key facts about the motion. Vicuna-13B [91] strikes a good balance of correctness, variety, and computation required. To obtain paraphrases, our prompt contains two parts. First the LLM-specific section that defines the requirement to provide helpful answers. For Vicuna, this begins: *A chat between....* This is followed by the paraphrasing prompt. We find the simple prompt: *Rephrase this video caption:* obtains varied, but factually correct paraphrases.

Verb-Variation Paraphrasing. Prompting in this way gives variety to vocabulary and sentence structure. However, the captions focus on the same low-level concepts as in our original caption. For instance, instead of *moves downwards quickly* we obtain *moves down rapidly* or *quickly descends*. Captions in downstream tasks are more high-level and do not describe each movement, instead describing this as an object is *dropped* or *falls*. Furthermore, many verbs have associations with specific objects *e.g.*, *drive* can be used to describe a car moving, while *fly* or *glide* is more suitable for a plane. We incorporate these ideas into our prompt to make our motion-focused representation more generalizable. Thus, our paraphrasing prompt becomes: *Rephrase this video caption using verbs related to o to describe the motion:*. The overall prompt is shown below.

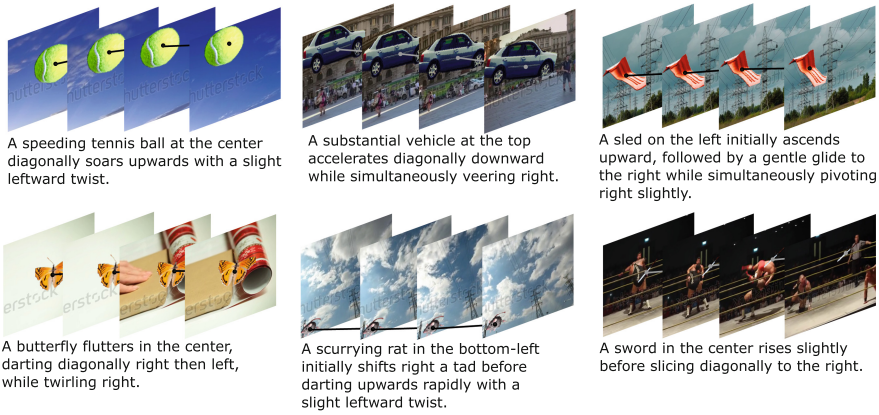

Fig. 5. Motion-Focused Video-Text Pairs. Our motion generation adds clearer and more varied motion to input videos. With our motion description and verb-variation paraphrasing, we are able to automatically create diverse descriptions of these motions.

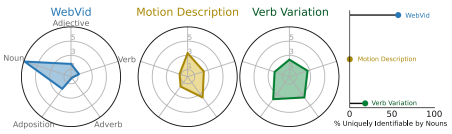

Fig. 6. Statistics of our captions. Our captions focus on different parts of speech and cannot be identified by nouns alone.

Fig. 7. Verbs in WebVid vs. our LocoMotion captions with verb variation. Our captions' verbs are more motion-relevant.

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
> USER: Hello!
> ASSISTANT: Hi!</s>
> USER: Rephrase this video caption using verbs related to o to describe the motion: t_{motion}
> ASSISTANT:

Taking all tokens generated after 'ASSISTANT: ' we obtain $t_{paraphrase}$ a rephrased version of t_{motion}. Figure 5 shows examples of the resulting video-text pairs $(v_{motion}, t_{paraphrase})$. Our captions focus more on adjectives, verbs, adverbs, adpositions than existing datasets as evidenced by the statistics in Fig. 6, and can rarely be identified by nouns alone. From Fig. 7 we also see our captions' verbs are more motion relevant than in WebVid, e.g. *dip*, *glide* or indicate temporal ordering, e.g. *change*, *begin*. While WebVid also contains different verbs, albeit much less, they are often focused on actions with specific objects, e.g. *cook*, *work*, or expressions, e.g. *smile*, *enjoy*.

4.4 Motion-Focused Pre-Training

Our goal is to learn a motion-focused representation $\mathcal{F}(v_{motion}, t_{paraphrase})$ which outputs the similarity of the contents of video v_{motion} to the corresponding description of the local object motion $t_{paraphrase}$. To achieve this we learn three functions, a vision encoder $\mathcal{F}_v$, a text encoder $\mathcal{F}_t$, and a multi-modal encoder $\mathcal{H}$. Thus $\mathcal{F}$ becomes:

$$\mathcal{F}(v, t) = \mathcal{H}(\mathcal{F}_v(v), \mathcal{F}_t(t)). \tag{6}$$

Our approach is agnostic to the choice of $\mathcal{F}_v, \mathcal{F}_t$ and $\mathcal{H}$ and can easily plug into existing video-language pre-training approaches. This can be done in two ways, depending on if the original pre-training has one or two stages. Multi-stage models [4,37] first train on image-text pairs before video-text pairs. We replace the second pre-training stage with our method, thereby removing the need for any non-synthetic video-text pairs. For single stage pre-training [10] we use $(v_{motion}, t_{paraphrase})$ and (v_{motion}, t) as our video-text pairs. To train $\mathcal{H}$, $\mathcal{F}_v$ and $\mathcal{F}_t$ we use three objectives: (i) Video-Text Contrast, aligns unimodel vision and text representations; (ii) Video-Text Matching, maximizes the score between matching video-text pairs; (iii) Video-Conditioned Masked Language Modelling, predicts masked text tokens from the video encoder. Since each object appears in different videos with different motions the contrastive and matching losses encourage the encoders to attend to the object's motion in addition to its appearance. The frequency of motion-relevant words in our caption also makes the video-conditioned masked language modeling loss key to learning motions.

5 Experiments

We first describe the implementation details and datasets used before ablating the components of our method and demonstrating it can be used in combination with different video-language pre-training models.

5.1 Implementation Details

Generating Motions. We pre-train using video frames of size $H \times W = 224 \times 224$. The object size is uniformly sampled from [32×32, 128×128]. We follow X-Paste [89] to create the set of objects O using Stable Diffusion [62] generating 60 samples for the 1203 categories in LVIS [26]. This results in ∼70k objects. For translation, we use $\delta \in [0, 10]$ displacement difference. Rotation angle θ_k is sampled from $[-25, 25]$. We use $K=3$ keyframes for translation and rotation.

Describing Motions. For each concept in the generated text we make the possible phrases equally likely. For instance, an object is considered small if it has a total area between 32×32 and 64×64 and big if it has a total area between 96×96 and 128×128. Full details are in the supplementary. For our verb-variation paraphrasing, we use Vicuna-13B v1.5 [91] with 16-bit floating point precision.

Video-Language Pre-training. Unless specified otherwise, we use Singularity-Temporal [37] as the backbone video-language model. Specifically, the visual

encoder $\mathcal{F}_v$ is BEiT$_{BASE}$ [5] pre-trained on ImageNet-21K [13], the text-encoder $\mathcal{F}_t$ uses the first 9 layers of BERT$_{BASE}$ [14] and the multimodal encoder $\mathcal{H}$ uses the last 3 layers of BERT$_{BASE}$ with the cross-attention randomly initialized. Since Singularity-Temporal has two stages of pre-training, we replace the second stage, *i.e.* the temporal stage, with LocoMotion using 4 video frames. In this stage, the model is optimized for 5 epochs using AdamW with an initial learning rate of 1e-4. The batch size is 20 per GPU and the model is trained on 8 A5000 NVIDIA GPUs. Details for VindLU [10] can be found in the supplementary.

Fine-tuning. Fine-tuning uses the same architecture as pre-training without the masked language modeling loss. We use an initial learning rate of 1e-5 with cosine decay to 1e-6. The model is fine-tuned using 4 A5000 NVIDIA GPUs each with a batch size of 16 for 10 epochs. During fine-tuning, we use 4 frames per video, while in testing we use 12. Code will be released on paper acceptance.

5.2 Datasets

WebVid. We use WebVid-2M [4] as the source of pre-training videos. This contains 2.5 million video clips scraped from the web alongside corresponding text descriptions of the videos. To make experiment time feasible, we use a smaller subset of WebVid for pre-training with 20% of the video-text pairs.

Something Something. We evaluate our model by fine-tuning on various motion-focused tasks. First, SSv2-Template and SSv2-Label [37], two text-to-video retrieval tasks based on Something Something v2 [24]. SSv2-Template uses 174 captions where the object is masked out, *e.g.* "Pushing [*something*] from left to right", while SSv2-Label uses the full label where the object replaces [*something*]. Both tasks use 168,913 videos for fine-tuning and 2,088 for testing. As one of the major benefits of pre-trained representations is the ability to easily adapt them, we also experiment with smaller subsets of SSv2-Template for fine-tuning: 10%, 20%, 25%, 33% and 50% of the fine-tuning data.

FineGym. As there is a lack of motion-focused video-language datasets, we borrow from unimodal video understanding and transform motion-heavy FineGym [63] into a video-language dataset. Captions are fine-grained *e.g.* "double salto backward tucked with 1 twist" and may only differ in a subtle aspect *e.g.* the direction ('backward'), position ('tucked') or the number of repetitions. We use both FineGym-99 and FineGym-288 subsets with 99 and 288 possible captions respectively. FineGym-99 has 20,484 videos while FineGym-288 has 22,959. For testing, we follow the setup of SSv2-Template [37] and sample up to 12 videos per label. This results in 1,188 videos for FineGym-99 and 3,036 for FineGym-288.

HumanML3D. We also increase the number of motion-focused video-language datasets by rendering motion-language dataset HumanML3D [25] into videos. The resulting data consists of 14,616 motions with 44,970 descriptions. We follow the data division from the original paper: 85% train 15% validation and 5% test.

Table 1. Model Components. All components contribute to a motion-focused representation. Paraphrasing is key to unlocking the potential of our motion description.

	R@1	R@5	R@10	Avg
Baseline	46.6	92.5	96.6	78.6
+ Generated Motion	52.3	90.2	96.0	79.5
+ Motion Description	55.2	92.5	97.7	81.8
+ Verb-Variation Paraphrasing	60.9	92.5	98.3	83.9

Table 2. Motion Description Phrases. The description of the translation motion $t_{translate}$ is key to our model's success.

	R@1	R@5	R@10	Avg
Original Caption	52.3	90.2	96.0	79.5
$t_{appearance}$	54.0	90.8	97.1	80.6
+ $t_{translate}$	56.9	93.7	98.3	83.0
+ t_{rotate}	56.9	93.1	97.7	82.6

Table 3. $t_{translate}$ Components. Direction of motion is most crucial, however all components contribute.

	R@1	R@5	R@10	Avg
$t_{appearance}$	54.0	90.8	97.1	80.6
+ speed	56.3	90.8	97.1	81.4
+ direction	58.6	92.5	96.6	82.6
+ distance	56.9	93.7	98.3	83.0

EPIC-Kitchens. Finally, we use EPIC-Kitchens-100 [11], a large-scale egocentric dataset of kitchen actions. We aim to retrieve the correct verb corresponding to the action, of which there are 97 possibilities. We use the standard splits which contains 67,217 video clips for fine-tuning and 9,668 video clips for validation.

For all datasets, we use the standard text-to-video retrieval metrics of R@1, R@5, R@10, and their average.

5.3 Ablation Study

To ablate the effectiveness of the individual components of our method we pre-train Singularity-Temporal [4] on the 20% subset of WebVid and fine-tune on the 25% subset of SSv2-Template to decrease training time.

Model Components. Table 1 demonstrates the contribution of each component of our motion-focused video-language pre-training. We first observe the notable benefits our overall model brings over the Singularity-Temporal baseline, trained with the original WebVid captions, *e.g.* +14.3 R@1. Adding the generated motion, without changing the video caption gives a small improvement (+0.9 average recall). While our generated motions can be unrealistic, their addition increases the presence and variety of motions. Replacing the original caption with our motion description results in a further improvement, *e.g.* +2.9 R@1. This causes the model to focus on the added motion and understand the language relevant to key aspects of motion such as the type, speed, and direction. However, generating motion descriptions with predefined phrases limits the variety and creates a domain gap between pre-training and downstream

Table 4. Verb-Variation Paraphrasing is key to the transferability of our learned representation as it allows more varied descriptions of the motions and the verbs used.

	R@1	R@5	R@10	Avg
No Paraphrasing	55.2	92.5	97.7	81.8
Basic Paraphrasing	58.1	92.0	98.9	83.0
Verb-Variation Paraphrasing	60.9	92.5	98.3	84.0

Table 5. Combination with Different Models (R@1). Our motion-focused pre-training can be effectively used in combination with different video-language models.

	SSv2 Temporal	SSv2 Label	FineGym 99	FineGym 288	Human ML3D	EPIC-Kitchens Verbs
Singularity-Temporal [37]						
Original	75.3	42.1	72.7	46.6	4.5	33.0
with LocoMotion	**76.4**	**43.5**	**74.8**	**51.0**	**5.2**	**34.0**
VindLU [10]						
Original	82.2	51.2	82.8	**58.9**	5.7	32.0
with LocoMotion	**84.5**	**53.1**	**84.9**	**58.9**	**6.5**	**39.2**

captions. Therefore, our verb-variation paraphrasing further unlocks the motion description potential, contributing an additional +5.7 R@1 improvement.

Motion Description Phrases. In Table 2 we investigate the effect of the caption components, keeping the generated motion the same. Surprisingly, replacing the original caption with $t_{appearance}$ increases the result. This is likely due to the concept of position that $t_{appearance}$ introduces. Even without describing position change, having awareness of position helps the model better understand motion in the downstream dataset. Adding the description of the generated motion with $t_{translate}$ contributes the largest increase in results, *e.g.* +2.9 R@1, as this includes the key motion concepts of speed, direction, and distance. Adding the description of the rotation motion gives little change to the result.

$t_{translate}$ Components. Since $t_{translate}$ is the main contributor to the increase in performance of our approach we further examine the contributions of its three components in Table 3. We see that speed, direction, and distance all contribute to the increase in performance. Direction is the biggest contributor (+1.2 average recall), as understanding the direction of motion is key to being able to distinguish between many different actions and activities.

Verb-Variation Paraphrasing. In Table 4 we investigate the effect of using the object name in the paraphrasing prompt to produce a greater variety of verbs. We compare this to a basic version without the object prompt where after the LLM-specific introduction, the prompt is instead *Rephrase this video*

caption:. From Table 4 we can see that indeed our verb-variation paraphrasing is more effective than basic paraphrasing.

5.4 Data-Efficient Fine-Tuning

A major benefit of large-scale pre-trained models is the ability to easily fine-tune and adapt them to different scenarios when labeled data is scarce. We investigate the benefit of our model for such scenarios in Fig. 8. Specifically, we fine-tune using different percentages of available data in SSv2-Template: 100%, 50%, 33%, 25%, 20%, and 10%. Our model outperforms the Singularity-Temporal baseline on all percentages of fine-tuning data. Our approach is particularly effective with limited data, for instance, with only 10% of the SSv2-Template fine-tuning data, we obtain a 13.3% absolute increase over Singularity-Temporal. This is due to the motion-focused representation our model has learned. It thus needs much less fine-tuning data to learn motion concepts required in downstream tasks.

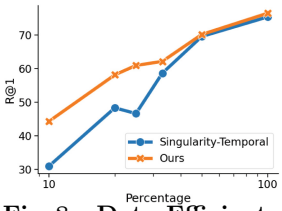

Fig. 8. Data-Efficient Fine-tuning. Our model is effective when fine-tuning with little data.

5.5 Combination with Different Models

Another benefit to our approach is that it can be combined with different video-language pre-training models. We demonstrate this in Table 5 by applying our approach to Singularity-Temporal [37] and VindLU [10] for various downstream tasks. We see our approach can be used effectively with both Singularity-Temporal and VindLU, improving results on all six downstream tasks.

Table 6. Comparison to Prior Works on SSv2-Template and SSv2-Label. Our approach outperforms prior works while pre-training with only 20% of the data.

	#Pre-train	SSv2-Template			SSv2-Label		
		R@1	R@5	R@10	R@1	R@5	R@10
Frozen [4]	5.0M	52.9	94.8	99.4	-	-	-
ALPRO [40]	5.0M	55.2	96.6	**100.0**	35.0	68.9	79.1
Lavender [43]	5.0M	67.8	98.3	**100.0**	46.9	77.2	84.5
Thoker *et al.* [71]	0.3M	69.0	98.3	99.4	31.0	61.1	73.3
CLIP4CLIP [48]	400.0M	77.0	96.6	98.3	43.1	71.4	80.7
Singularity-T [37]	5.0M	77.0	98.9	99.4	44.1	73.5	82.2
VindLU [10]	5.0M	82.2	98.9	-	51.2	78.8	-
LocoMotion	1.2M	**84.5**	**99.4**	99.4	**53.1**	**80.2**	**87.0**

5.6 Comparative Results

We compare to prior works on SSv2-Template and SSv2-Label in Table 6. We compare with fine-tuned video-text retrieval models [4,10,37,40,48] and Lavender [43] as well as the motion-focused video-only model from Thoker *et al.* [71]. Our approach outperforms prior works even when trained with only 20% of the available videos. We also compare our approach on ActionBench [77] (Table 7) which probes a model's temporal understanding with two video-text tasks: distinguishing an action vs. its antonym (AA) and distinguishing a video from its reserved counterpart (VR). On both tasks, we outperform all models tested in [77]. Our model even outperforms Singularity with the DVDM losses specialized for these tasks [77].

Table 7. ActionBench

	AA	VR
InternVid [76]	51.8	48.3
Clip-ViP [82]	70.2	53.6
Singularity [37]	48.9	48.9
+ DVDM losses	82.4	68.8
LocoMotion	**83.6**	**80.7**

6 Conclusion

This paper highlights the spatial focus of video language representations, caused by the caption content of current datasets, and instead learns motion-focused representations. We present LocoMotion, a video-language pre-training approach. We learn from augmented videos with synthetic local object motion and corresponding motion descriptions which include various motion-relevant concepts. By prompting LLMs to paraphrase motion descriptions using verbs relevant to the moving object, we increase the variety in captions and make better links between low-level motion primitives and high-level verbs. Although our generated motions can be unrealistic, experiments show our method is effective for motion-focused tasks, particularly when limited fine-tuning data is available.

Acknowledgements. This work is supported by the Dutch Research Council (NWO) under a Veni grant (VI.Veni.222.160) and KAUST Center of Excellence for Generative AI under award number 5940.

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5803–5812 (2017)
2. Bagad, P., Tapaswi, M., Snoek, C.G.M.: Test of time: Instilling video-language models with a sense of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2503–2516 (2023)
3. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2020)
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). pp. 1728–1738 (2021)
5. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: International Conference on Learning Representations (ICLR) (2021)

6. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the" video" in video-language understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2917–2927 (2022)
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–970 (2015)
8. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
9. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18000–18010 (2023)
10. Cheng, F., Wang, X., Lei, J., Crandall, D., Bansal, M., Bertasius, G.: Vindlu: A recipe for effective video-and-language pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10739–10750 (2023)
11. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision (IJCV) pp. 1–23 (2022)
12. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: Tclr: Temporal contrastive learning for video representation. Computer Vision and Image Understanding (CVIU) **219**, 103406 (2022)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2019)
15. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H.: Motion-aware contrastive video representation learning via foreground-background merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
16. Doughty, H., Laptev, I., Mayol-Cuevas, W., Damen, D.: Action modifiers: Learning from adverbs in instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
17. Doughty, H., Snoek, C.G.M.: How do you do it? fine-grained action understanding with pseudo-adverbs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
18. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
19. Feichtenhofer, C., Fan, H., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
20. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

21. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
22. Gavrilyuk, K., Jain, M., Karmanov, I., Snoek, C.G.M.: Motion-augmented self-training for video recognition at smaller scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
23. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). pp. 1396–1406 (2021)
24. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
25. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (2022)
26. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
27. Han, T., Xie, W., Zisserman, A.: Self-supervised co-training for video representation learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
28. Hong, J., Fisher, M., Gharbi, M., Fatahalian, K.: Video pose distillation for few-shot, fine-grained sports action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
29. Hu, H., Dong, S., Zhao, Y., Lian, D., Li, Z., Gao, S.: Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
30. Huang, D., Wu, W., Hu, W., Liu, X., He, D., Wu, Z., Wu, X., Tan, M., Ding, E.: Ascnet: Self-supervised video representation learning with appearance-speed consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
31. Huang, Z., Zhang, S., Jiang, J., Tang, M., Jin, R., Ang, M.H.: Self-supervised motion learning from static images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
32. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2758–2766 (2017)
33. Jenni, S., Jin, H.: Time-equivariant contrastive video representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
34. Kim, M., Kwon, H., Wang, C., Kwak, S., Cho, M.: Relational self-attention: What's missing in attention for video understanding. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
35. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 706–715 (2017)

36. Kwon, H., Kim, M., Kwak, S., Cho, M.: Learning self-similarity in space and time as generalized motion for video action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
37. Lei, J., Berg, T.L., Bansal, M.: Revealing single frame bias for video-and-language learning. In: Proceedings of the Annual Meeting for the Association of Computational Linguistics (ACL) (2023)
38. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7331–7341 (2021)
39. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)
40. Li, D., Li, J., Li, H., Niebles, J.C., Hoi, S.C.: Align and prompt: Video-and-language pre-training with entity prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4953–4963 (2022)
41. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
42. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)
43. Li, L., Gan, Z., Lin, K., Lin, C.C., Liu, Z., Liu, C., Wang, L.: Lavender: Unifying video-language understanding as masked language modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23119–23129 (2023)
44. Li, T., Foo, L.G., Ke, Q., Rahmani, H., Wang, A., Wang, J., Liu, J.: Dynamic spatio-temporal specialization learning for fine-grained action recognition. In: European Conference on Computer Vision (ECCV) (2022)
45. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
46. Lin, Z., Qi, S., Zhengyang, S., Changhu, W.: Inter-intra variant dual representations for self-supervised video recognition. In: British Machine Vision Conference (BMVC) (2021)
47. Liu, X., Li, Y.L., Zeng, A., Zhou, Z., You, Y., Lu, C.: Bridging the gap between human motion and action semantics via kinematic phrases. In: European Conference on Computer Vision (ECCV) (2024)
48. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
49. Mac, K.N.C., Joshi, D., Yeh, R.A., Xiong, J., Feris, R.S., Do, M.N.: Learning motion in feature space: Locally-consistent deformable convolution networks for fine-grained action detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
50. Mavroudi, E., Bhaskara, D., Sefati, S., Ali, H., Vidal, R.: End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) (2018)

51. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2630–2640 (2019)
52. Moltisanti, D., Keller, F., Bilen, H., Sevilla-Lara, L.: Learning action changes by measuring verb-adverb textual relationships. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23110–23118 (2023)
53. Ni, B., Paramathayalan, V.R., Moulin, P.: Multiple granularity analysis for fine-grained action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
54. Ni, J., Zhou, N., Qin, J., Wu, Q., Liu, J., Li, B., Huang, D.: Motion sensitive contrastive learning for self-supervised video representation. In: European Conference on Computer Vision (ECCV) (2022)
55. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
56. Peihao, C., Deng, H., Dongliang, H., Xiang, L., Runhao, Z., Shilei, W., Mingkui, T., Chuang, G.: Rspnet: Relative speed perception for unsupervised video representation learning. In: The AAAI Conference on Artificial Intelligence (AAAI) (2021)
57. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV). pp. 480–497. Springer (2022)
58. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
59. Piergiovanni, A., Ryoo, M.S.: Fine-grained activity recognition in baseball videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018)
60. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
61. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
62. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
63. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
64. Shao, D., Zhao, Y., Dai, B., Lin, D.: Intra-and inter-action understanding via temporal action parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
65. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In: European Conference on Computer Vision (ECCV) (2016)

66. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
67. Sun, B., Ye, X., Yan, T., Wang, Z., Li, H., Wang, Z.: Fine-grained action recognition with robust motion representation decoupling and concentration. In: Proceedings of the ACM International Conference on Multimedia (ACMMM) (2022)
68. Tao, L., Wang, X., Yamasaki, T.: Pretext-contrastive learning: Toward good practices in self-supervised video representation leaning. arXiv preprint arXiv:2010.15464 (2021)
69. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: International Conference on Learning Representations (ICLR) (2023)
70. Thoker, F.M., Doughty, H., Bagad, P., Snoek, C.G.M.: How severe is benchmark-sensitivity in video self-supervised learning? In: European Conference on Computer Vision (ECCV) (2022)
71. Thoker, F.M., Doughty, H., Snoek, C.G.M.: Tubelet-contrastive self-supervision for video-efficient generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13812–13823 (2023)
72. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
73. Wang, G., Zhou, Y., Luo, C., Xie, W., Zeng, W., Xiong, Z.: Unsupervised visual representation learning by tracking patches in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
74. Wang, J., Gao, Y., Li, K., Lin, Y., Ma, A.J., Cheng, H., Peng, P., Huang, F., Ji, R., Sun, X.: Removing the background by adding the background: Towards background robust self-supervised video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
75. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4581–4591 (2019)
76. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: International Conference on Learning Representations (ICLR) (2024)
77. Wang, Z., Blume, A., Li, S., Liu, G., Cho, J., Tang, Z., Bansal, M., Ji, H.: Paxion: Patching action knowledge in video-language foundation models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
78. Xiao, F., Tighe, J., Modolo, D.: Maclr: Motion-aware contrastive learning of representations for videos. In: European Conference on Computer Vision (ECCV) (2022)
79. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the ACM International Conference on Multimedia (ACMMM). pp. 1645–1653 (2017)
80. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot

video-text understanding. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021)
81. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5288–5296 (2016)
82. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In: International Conference on Learning Representations (ICLR) (2023)
83. Yang, C., Xu, Y., Shi, J., Dai, B., Zhou, B.: Temporal pyramid network for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
84. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: Merlot: Multimodal neural script knowledge models. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 23634–23651 (2021)
85. Zhang, C., Gupta, A., Zisserman, A.: Temporal query networks for fine-grained video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
86. Zhang, H., Xu, X., Han, G., He, S.: Context-aware and scale-insensitive temporal repetition counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
87. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
88. Zhang, Y., Shao, L., Snoek, C.G.M.: Repetitive activity counting by sight and sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
89. Zhao, H., Sheng, D., Bao, J., Chen, D., Chen, D., Wen, F., Yuan, L., Liu, C., Zhou, W., Chu, Q., et al.: X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In: International Conference on Learning Representations (ICLR) (2023)
90. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. International Conference on Machine Learning (ICML) (2024)
91. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
92. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8746–8755 (2020)