

# Action100M: A Large-scale Video Action Dataset

Delong Chen<sup>1,2</sup> Tejaswi Kasarla<sup>1,3</sup> Yejin Bang<sup>1</sup> Mustafa Shukor<sup>1,4</sup> Willy Chung<sup>1,4</sup>  
 Jade Yu<sup>1</sup> Allen Bolourchi<sup>1</sup> Théo Moutakanni<sup>1</sup> Pascale Fung<sup>1,2</sup>

<sup>1</sup>Meta FAIR <sup>2</sup>HKUST <sup>3</sup>University of Amsterdam <sup>4</sup>Sorbonne Université

## Abstract

*Inferring physical actions from visual observations is a fundamental capability for advancing machine intelligence in the physical world. Achieving this requires large-scale, open-vocabulary video action datasets that span broad domains. We introduce ACTION100M, a large-scale dataset constructed from 1.2M Internet instructional videos (14.6 years of duration), yielding  $O(100)$  million temporally localized segments with open-vocabulary action supervision and rich captions. ACTION100M is generated by a fully automated pipeline that (i) performs hierarchical temporal segmentation using V-JEPA 2 embeddings, (ii) produces multi-level frame and segment captions organized as a TREE-OF-CAPTIONS, and (iii) aggregates evidence with a reasoning model (GPT-OSS-120B) under a multi-round SELF-REFINE procedure to output structured annotations (brief/detailed action, actor, brief/detailed caption). Training VL-JEPA on ACTION100M demonstrates consistent data-scaling improvements and strong zero-shot performance across diverse action recognition benchmarks, establishing Action100M as a new foundation for scalable research in video understanding and world modeling.*

## 1. Introduction

Making machine intelligence useful in the physical world requires AI models that not only understand world states (e.g., objects and their attributes), but also recognize **physical actions** that interact with the world and induce state transitions. Powered by supervision from large datasets [6, 29, 37, 39], world state perception in frontier models has advanced rapidly [3, 7, 19, 21, 31, 32]. In contrast, the capability of understanding *actions* is comparably limited, largely due to the absence of robust data foundations. Existing video action datasets are mainly separately developed for individual narrow domains (e.g., cooking, toy assembly), and remain in limited scale (e.g., less than 1 million action instances).

Advancement in data foundation will enable the develop-

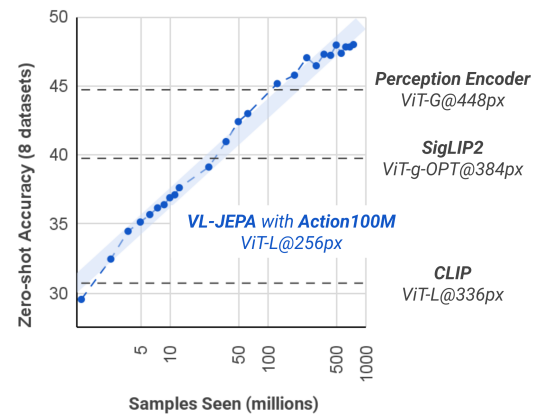


Figure 1. Scaling on ACTION100M improves zero-shot action recognition consistently.

ment of **open-domain** and **open-vocabulary** video action recognizers. These models will support embodied learning, facilitate wearable assistive applications, and advance physical world modeling [1, 8, 16, 17, 23, 36, 41, 43, 52]. Achieving this necessitates training data that is sufficiently *large scale*, maintains *high quality*, and spans *broad diversity*.

To address these challenges, we introduce ACTION100M, a large-scale video action dataset containing  $O(100)$  million action instances, annotated via a fully automated pipeline. The construction process begins with online instructional videos [33], which capture people interacting with the physical world in diverse activities. Dense action labels are produced using a family of frontier open-source models, including V-JEPA 2 [5], PerceptionLM [20], Llama-3.2-Vision, and GPT-OSS [2]. The TREE-OF-CAPTIONS [15] and SELF-REFINE mechanisms [17] are leveraged to reduce hallucinations and ensure annotation quality. This pipeline yields a hierarchy of temporal segments annotated with structured fields, including brief/detailed action descriptions, the actor, and brief/detailed video captions.

The annotation pipeline was executed on 1.2 million YouTube videos (14.6 years of duration), spending approx-

imately 1.3 million V100 GPU hours for segmentation and captioning, and 0.3 million H100/H200 GPU hours for LLM aggregation. It derives 147 million segment-level annotations, totaling 21.3 billion English words.

To demonstrate the utility of ACTION100M, we train VL-JEPA models [18] on the dataset. VL-JEPA is a vision-language model that extracts visual embeddings with V-JEPA 2 [5] encoder and predicts target embedding generated by a text encoder. The model uses InfoNCE loss to learn an aligned embedding space that allows CLIP-style open-vocabulary classification [35]. Evaluation is conducted on eight downstream video action benchmarks, including Something-something-v2 [25], EPIC-KITCHENS-100 [22], EgoExo4D Keysteps [26], Kinetics-400 [27], COIN [42], and CrossTask [59]. These benchmarks evaluate actions spanning diverse domains and different levels of abstraction, from low-level fine-grained ones like “turn on light”, “cut tomato”, to high-level procedural ones like “make Irish coffee”, “assemble computer”, etc.

Fig. 1 demonstrates the **effectiveness of scaling** for zero-shot action recognition: VL-JEPA performance improves consistently as the amount of ACTION100M training data increases (x-axis shows effective batch size  $\times$  number of iterations). Our VL-JEPA (ViT-L at 8 frames per input in 256px) outperform CLIP [35], SigLIP2 [45], and Perception Encoder [11] which have seen much more samples (13B, 40B, and 86B, respectively, while VL-JEPA only 3B) and use larger backbone and higher input resolution. At the same time, the caption annotations in ACTION100M allows VL-JEPA to establish robust vision-language alignment that allows effective zero-shot text-to-video retrieval. Compared to CLIP, SigLIP2, and Perception Encoder, VL-JEPA achieves higher average retrieval recall@1 on eight benchmarks: MSR-VTT [54], ActivityNet [12], DiDeMo [4], MSVD [14], YouCook2 [58], PVDBench [11], Dream-1K [48], and VDC-1K [13].

## 2. Related Works

We organize prior efforts into two main categories: video action datasets, which focus on annotated action segments, and video captioning datasets, which provide descriptive text for video content. We present a comparison of ACTION100M with existing action and caption datasets in Appendix Table ??.

**Video Action Datasets.** Standard video action datasets, such as COIN [42], CrossTask [59], and YouCook2 [58], have played a pivotal role in advancing action detection in untrimmed videos. These datasets typically segment long activities into semantically meaningful steps, each paired with textual descriptions, and are primarily composed of instructional videos covering procedural tasks. The collection methodologies for video datasets vary significantly:

- Participant-recorded datasets (e.g., Breakfast [28], As-

sembly101 [38]) provide dense, hierarchical annotations in controlled environments with high-quality labels through manual data collection.

- Internet-mined datasets (e.g., CrossTask, COIN) leverage external taxonomies to achieve broader coverage of activities. By mining videos from online sources, these datasets capture a wider range of human actions.
- Egocentric datasets (e.g., EgoProceL [9], Ego4D Goal-Step [40]) introduce data-driven, hierarchical taxonomies and dense step annotations, particularly in domains such as cooking. These datasets support research on goal inference and long-term temporal modeling by capturing first-person perspectives and fine-grained activity sequences.

Despite these advances, existing datasets are often constrained by manual annotation bottlenecks, limited domain coverage, or lack of scale.

While datasets such as COIN [42] and YouCook2 [58] provide valuable annotated instructional videos, they are constrained by manual annotation and limited activity coverage. Internet-mined datasets expand diversity but often lack fine-grained, hierarchical labels. Egocentric datasets introduce new perspectives but remain domain-specific. ACTION100M marks a step forward in this landscape.

In contrast, ACTION100M leverages large-scale online instructional videos [33] to deliver over 100 million action instances with multi-level, open-vocabulary annotations. This enables robust research in open-domain action recognition and world modeling, as summarized in Table ???. By addressing annotation bottlenecks and supporting fine-grained temporal reasoning, ACTION100M provides the way for future advances in video understanding.

**Video Caption Datasets.** A number of large-scale video-language datasets have been proposed in recent years. InternVid [51], HD-VILA-100M [55], and VidChapters-7M [57] collect millions of videos or hundreds of millions of clips with associated textual descriptions. However, the majority of these captions are obtained from noisy sources such as automatic speech recognition (ASR) transcripts or video metadata. As a result, the captions are typically short, generic, and only weakly aligned with the underlying physical actions, limiting their usefulness for learning fine-grained action representations or detailed world models.

To address the limitations of ASR-only captions, recent instruction-style datasets leverage powerful vision-language models to produce richer descriptions. OpenVid-1M [34], UltraVideo [56], and VideoUFO [50] generate longer, more detailed, and instruction-like captions that better capture object interactions and step-by-step activities. Nonetheless, these datasets are typically built on a relatively small number of source videos compared to the largest ASR-based corpora, which constrains their coverage of diverse environments and long-horizon activities.

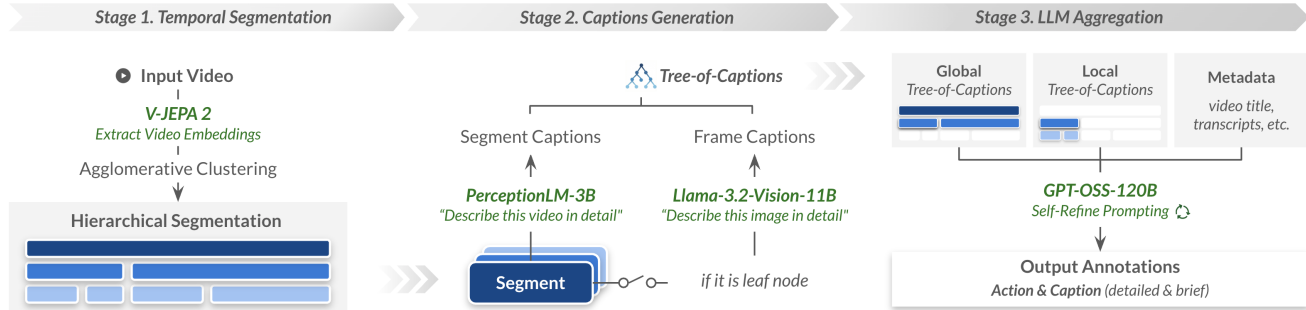


Figure 2. **ACTION100M Data Pipeline.** Our pipeline first applies hierarchical temporal segmentation to decompose the video into semantically coherent segments at multiple temporal scales. For each segment, we generate video caption and frame captions, capturing both temporal and spatial information. Next, we prompt LLM to aggregate the captions, extracting final annotations.

Koala-36M [49] further scales LLM-captioned video data, but is primarily oriented toward video generation rather than structured action understanding.

For video understanding and perception-centric applications, PE Video and PLM-Video-Auto [20] combine millions of videos with LLM-generated captions. While they improve caption quality over pure ASR pipelines, they generally assign a single caption per segment and do not explicitly encode temporal hierarchies within videos. Consequently, they provide limited supervision for modeling multi-scale structure in activities (e.g., steps, sub-tasks, and overarching tasks), which is crucial for learning world models that reason over extended sequences of actions and their context.

Moreover, many existing video captioning datasets rely heavily on manual or semi-manual annotation, which is expensive and restricts scale, or they prioritize scale at the cost of fine-grained, action-centric detail and temporal structure. In contrast, ACTION100M is constructed to combine the scale of ASR-driven corpora with the descriptiveness of recent LLM-captioned datasets. It introduces a Tree-of-Captions structure derived from hierarchical temporal segmentation, providing multiple levels of captions that capture both fine-grained actions and broader contextual narratives. This design preserves temporal hierarchy in a dense-captioning-like manner, making ACTION100M particularly suited for training large-scale world models and action-centric video understanding systems.

### 3. Action100M Data Pipeline

This section describes an automated, scalable pipeline for constructing ACTION100M. The pipeline extends and improves the data generation procedure introduced in Vision Language World Model [17]. Fig. 2 provides an overview of the pipeline, and Fig. 3 visualizes an example of annotation. The pipeline has three stages:

1. We decompose each video into a hierarchy of tempo-

rally coherent segments, which captures both short, fine-grained motions and long, procedural steps, enabling action supervision at multiple levels of abstraction.

2. For every segment, we generate complementary frame- and video-level captions and organize them into a TREE-OF-CAPTIONS. Captioning is a task for which modern VLMs are extensively trained, and even medium-sized models can produce reliable results.
3. We prompt an LLM to aggregate information from the TREE-OF-CAPTIONS and to extract structured annotations. Assembling evidence across multiple caption levels provably reduces hallucinations [15].

Overall, rather than applying heavy VLMs directly to entire videos, our pipeline first converts videos into hierarchical, text-based representations (TREE-OF-CAPTIONS, inspired by Pyramid-of-Captions [15]) and only leverages strong reasoning models in the final stage, which is purely text-based. This design allow us to obtain reliable annotations while keeping the overall computation cost manageable. In the following, we describe each stage in detail.

**Stage 1. Temporal Segmentation.** Each video is first transformed into a temporally dense sequence of visual embeddings using the V-JEPA 2 encoder. Frames are uniformly sampled at one out of every four raw frames to approximate the temporal resolution used during V-JEPA 2 pretraining. The video is then divided into overlapping temporal windows, each containing 64 sampled frames, with an eight-frame stride between consecutive windows. Each window is independently processed by the V-JEPA 2 ViT-g-384 encoder, which outputs a sequence of spatial-temporal visual tokens. We perform spatial average pooling, resulting in a per-frame feature of dimensionality equal to the encoder’s hidden size. Because adjacent windows overlap, multiple representations are produced for shared frames. These are accumulated and averaged to form a single, temporally consistent embedding per frame across the entire video.

To capture actions across multiple temporal scales, we

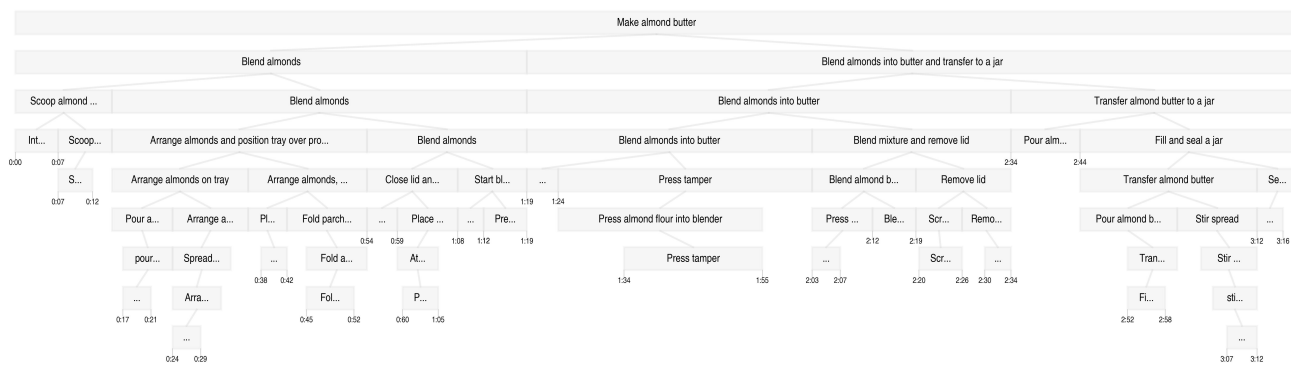


Figure 3. **Example of hierarchical structure in Action100M annotations** (with brief action description labels shown). Source video: [url](#). **Brief caption** of the entire video: *A woman roasts almonds, blends them into butter, and pours the butter into a jar.* **Detailed caption:** *The video opens with the presenter in a bright kitchen speaking to the camera. She spreads raw almonds on a parchment-lined tray, places the tray in a pre-heated 350 °F oven, and after roasting lets the nuts cool to room temperature. She then transfers the almonds to a Vitamix blender, removes the lid, inserts a tamper, and sets the machine on high. While the blender runs she presses the almonds down with the tamper, first creating a fine flour and then a thick creamy butter within about one minute. She pours the almond butter into a clear storage jar, scoops it with a large wooden spoon and stirs it to smooth the surface, then concludes the segment with a brief thank-you.*

apply *hierarchical agglomerative clustering* to the sequence of frame-level representations. A local temporal connectivity constraint links each frame only to its immediate neighbors, ensuring that merges occur only between contiguous time spans. Clustering proceeds bottom-up using Ward linkage, which minimizes the variance within each segment at every merge step. Starting from individual frames or short windows, adjacent segments are iteratively merged to minimize intra-cluster variance. This process produces a hierarchical tree decomposition of the video, where each node corresponds to a contiguous, semantically coherent temporal segment. Lower levels of the hierarchy correspond to fine-grained atomic motions, while higher levels capture coarser activities. We retain only nodes whose duration is larger than 0.5 seconds, ensuring that each segment is semantically meaningful.

**Stage 2. Caption Generation.** After hierarchical segmentation, we annotate each node in the video tree with captions, ensuring that both local fine-grained and global contextual information are captured within the same representation. The captioning process operates in two complementary modes: mid-frame captioning for fine-grained spatial details and video-segment captioning for temporal dynamics.

For every leaf node (representing the smallest contiguous action segment), we extract a key frame at the midpoint of its temporal span. These mid-frames are processed using Llama-3.2-Vision-11B, prompted with “Describe this image in detail.” to generate frame captions. For higher-level nodes representing longer temporal spans, we apply Perception-LM-3B. Each segment is sampled into 32 evenly spaced frames between its start and end times

at  $320^2$  resolution, and the model is prompted with “Describe this video in detail.” to generate segment-level captions. Both models are configured with a generation limit of 1024 tokens and can be run on a single NVIDIA V100 32GB GPU.

**Stage 3. LLM Aggregation.** For each node in the TREE-OF-CAPTIONS, we generate structured action annotations by prompting GPT-OSS-120B to extract five fields of information: brief action description, detailed action description, actor, brief video caption, and detailed video caption. This extraction is performed through three iterative rounds of SELF-REFINE to improve consistency and quality.

The inputs to the LLM include the current node’s caption, its children’s captions formatted in depth-first order, as well as global context such as root captions (within a limited depth) and video metadata (including title, description, and ASR transcript). Each node in the TREE-OF-CAPTIONS is processed independently. Nodes with a duration shorter than four seconds are discarded. For the remaining nodes, we query GPT-OSS-120B to infer clean, structured textual representations that unify and denoise information from multiple caption sources. More details about the process and the prompt to the GPT-OSS-120B are in the supplementary.

## 4. Dataset Analysis

**Source Videos.** Action100M is built on 1,199,096 face-blurred videos sourced from HowTo100M [33], corresponding to a total video duration of approximately 14.6 years. Since the original HowTo100M dataset was released



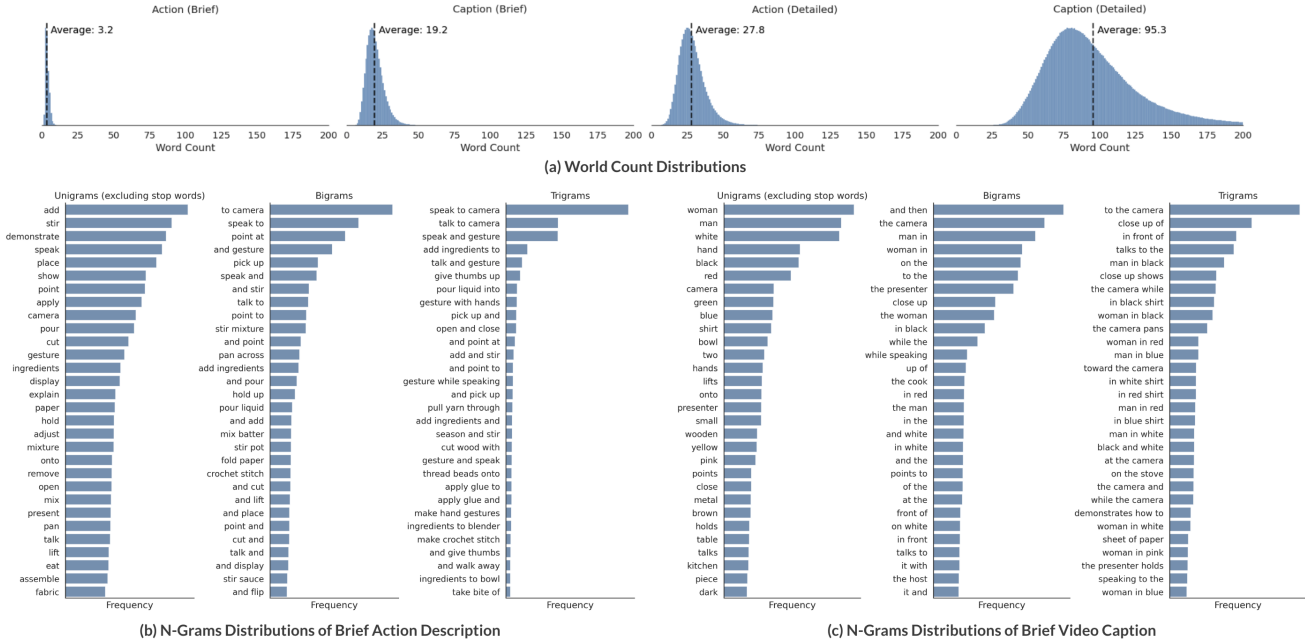


Figure 6. **Statistics of generated textual annotations in Action100M.** (a) Word-count distributions for four annotation types: brief action, brief caption, detailed action, and detailed caption. Dashed lines indicate the mean length of each annotation type. (b-c) Top unigrams, bigrams, and trigrams (excluding stop words) in brief action descriptions and brief video captions.



Figure 7. **Sunburst visualizations of frequent action compositions.** Each sunburst shows the hierarchical co-occurrence structure centered on one of the five most frequent verbs in brief action descriptions.

frame per input. We use DataComp-1B [24], mixing original caption and model generated caption from [30], and YFCC-100M [44] captioned by PaliGemma [10] and Florence-2 [53].

- In **Stage 2**, we process video data with eight frames per input. We continue from stage 1 checkpoint and use ACTION100M data (mixing all four field, with detailed action and caption downsampled by a half), and PerceptionLM-3B labelled action description and video caption for each ACTION100M segment.
- In **Stage 3**, we increase the number of frames per input to 32, and unfreeze the V-JEPA 2 encoder. This leads to higher CUDA memory consumption and lower batch size, and we use gradient accumulation of 4 steps and lower learning rate (1e-5 instead of 5e-5).

We benchmark resulting models with existing founda-

tion models, including CLIP [35], SigLIP2 [45], and Perception Encoder [11]. We evaluate on two tasks:

- **Zero-shot action recognition** (top-1 accuracy) on eight benchmarks: Something-something-v2 (SSv2) [25], EPIC-KITCHENS-100 (EK-100) [22], EgoExo4D Keysteps [26], Kinetics-400 [27], COIN [42], and CrossTask [59]. For COIN and Crosstask, we evaluate both segment-level step recognition and global task recognition.
- **Zero-shot text-to-video retrieval** (recall@1) on eight benchmarks: MSR-VTT [54], ActivityNet [12], DiDeMo [4], MSVD [14], YouCook2 [58], PVD-Bench [11], Dream-1K [48], and VDC-1K [13].

Table 1. Details of VL-JEPA’s three-stage training procedure. Stage 3 uses a gradient accumulation of 4 steps.

Training Stage	Vision encoder	#Frames per input	Training Data	Batch Size	#Iterations
Stage 1	Frozen	1	Image-text data (DataComp-1B, YFCC-100M)	24,576	100k
Stage 2	Frozen	8	Action100M	12,288	60k
Stage 3	Unfrozen	32	Action100M	3,072×4	10k

Table 2. **Main results.** We evaluate zero-shot performance on eight action recognition dataset. and eight video retrieval datasets.

Model	#Parameters	#Samples Seen	#Frames	Action Recognition (Top-1 Accuracy)										Text-to-video Retrieval (Recall@1)									
				Average	SSv2	EK100	EgoExo4D	Kinetics-400	COIN (SR)	COIN (TR)	CrossTask (SR)	CrossTask (TR)	Average	MSR-VTT	ActivityNet	DiDeMo	MSVD	YouTube2	PVD-Bench	Dream-1k	VDC-1k		
CLIP	RN50 (224px)	75M		21.8	2.1	1.5	2.1	41.4	8.6	39.0	10.9	68.7	28.3	28.7	17.7	24.7	29.7	5.1	27.6	47.2	46.0		
	ViT-B (224px)	124M	12.8B	25.4	3.1	1.3	2.8	49.5	11.2	47.3	16.2	71.5	29.3	31.0	19.5	25.7	34.0	6.1	27.0	48.5	42.9		
	ViT-L (336px)	389M		30.7	3.8	3.7	2.6	58.3	14.7	63.5	20.8	78.5	35.3	35.9	23.4	30.7	41.9	7.9	36.7	56.8	49.3		
SigLIP2	ViT-B (224px)	375M		33.9	5.2	2.3	4.5	57.8	20.6	69.9	27.7	82.9	39.6	40.2	25.0	32.1	48.6	13.8	52.1	60.9	43.7		
	ViT-L (384px)	882M	40B	38.6	5.9	4.5	6.4	63.6	24.2	78.5	35.1	90.8	45.4	41.6	32.7	35.1	53.5	19.0	59.2	71.6	50.9		
	ViT-g (384px)	1.9B		39.8	6.1	6.1	5.6	68.0	26.0	80.4	35.1	90.8	47.5	43.4	33.9	38.9	56.0	22.2	60.4	73.0	52.5		
PE-Core	ViT-B (224px)	448M	58B	37.2	5.8	3.3	6.0	65.4	21.5	77.1	26.9	91.8	44.9	46.5	35.4	35.3	49.1	15.2	59.8	68.7	49.2		
	ViT-L (384px)	671M	58B	42.9	9.3	6.0	11.6	73.4	27.1	83.3	37.5	95.3	50.2	48.9	41.7	40.8	56.2	22.5	64.7	75.9	51.0		
	ViT-G (448px)	2.3B	86B	44.7	9.0	6.4	13.6	<b>76.4</b>	29.0	<b>86.0</b>	40.3	<b>97.2</b>	58.1	<b>51.6</b>	49.1	44.5	<b>58.7</b>	26.0	77.0	89.2	68.5		
VL-JEPA	Stage 1		2.4B	8	21.5	3.9	1.0	3.2	39.5	12.1	40.6	20.9	50.9	32.6	27.7	20.9	26.6	34.2	3.4	51.6	52.9	43.8	
	Stage 2	ViT-L (256px)	1.6B	3.1B	8	48.0	18.4	16.8	25.7	61.2	43.3	72.7	62.0	83.9	59.5	37.1	54.2	46.0	45.7	33.9	80.2	91.5	87.3
	Stage 3		3.3B	32	<b>52.5</b>	<b>19.3</b>	<b>21.8</b>	<b>33.2</b>	64.8	<b>47.4</b>	79.4	<b>64.5</b>	89.6	<b>63.7</b>	40.0	<b>64.9</b>	<b>50.0</b>	49.0	<b>40.4</b>	<b>83.1</b>	<b>93.3</b>	<b>88.8</b>	

## 5.2. Main Results

As shown in Tab. 2, with ACTION100M, VL-JEPA achieves higher average action recognition performance and average video retrieval performance. Thanks to strong V-JEPA 2 encoder and dense fine-grained action annotation in ACTION100M, VL-JEPA is particularly strong on **motion-focused** tasks, such as Something-something-v2, EPIC-KITCHENS-100, EgoExo4D Keysteps, step recognition on COIN and CrossTask.

Fig. 8 visualize evolution of per-dataset accuracy in each stages, with x-axis being log-scale number of samples seen in each stage. We see that scaling yields consistent improvement, especially on motion-focused datasets mentioned earlier. We also see a significant jump from stage 1 to stage 2, indicating that image-only training alone is insufficient for action recognition.

## 5.3. Effectiveness of ACTION100M Pipeline

We compare the effectiveness of different data under a controlled training setting. All models are initialized from the same stage 1 checkpoint, and further trained for 20k steps (20.48M samples seen). Fig. 9 visualizes the resulting performance improvement compared to the stage 1 initialization.

We see that brief action descriptions in ACTION100M are highly effective in terms of improving zero-shot action recognition. It outperform the direct PLM-3B pseudo labeling baseline, showing the effectiveness of hierarchical captioning and LLM aggregation. Detailed captions in ACTION100M also show advantages over PLM-Video-Auto

captions on most benchmarks. At the same time, ACTION100M is much larger than it (100M vs 3M). Training with Ego4D atomic action description significantly improve egocentric action recognition performance on EK-100 and EgoExo4D, while not being effective on other domains.

## 5.4. Effectiveness of Semantic Resampling

A major challenge in large-scale action datasets is the long-tailed distribution of action frequencies, which can bias model training. To address this, we explore semantic resampling inspired by DINOv2 [47] to promote a more balanced sampling of actions.

We begin by embedding all brief action descriptions from the TREE-OF-CAPTIONS using EmbeddingGemma-300M [46]. To address redundancy, we deduplicate these actions by hashing their text, ensuring that repeated descriptions are consolidated. Following deduplication, we apply k-means clustering to the resulting embeddings, with cluster counts of  $k = \{10^3, 10^4, 10^5\}$ . This clustering groups semantically similar actions together, allowing us to control the granularity of the action space. From each cluster, actions are sampled uniformly with replacement until the target dataset size is reached, which ensures that both frequent and rare actions are adequately represented in the training data.

To empirically validate the effectiveness of this pipeline, we curated a 10M-action dataset using semantic resampling across all tested values of  $k$ . All models were initialized from the same stage 1 checkpoint and subsequently trained for 10k steps, corresponding to 10.24M samples seen. We

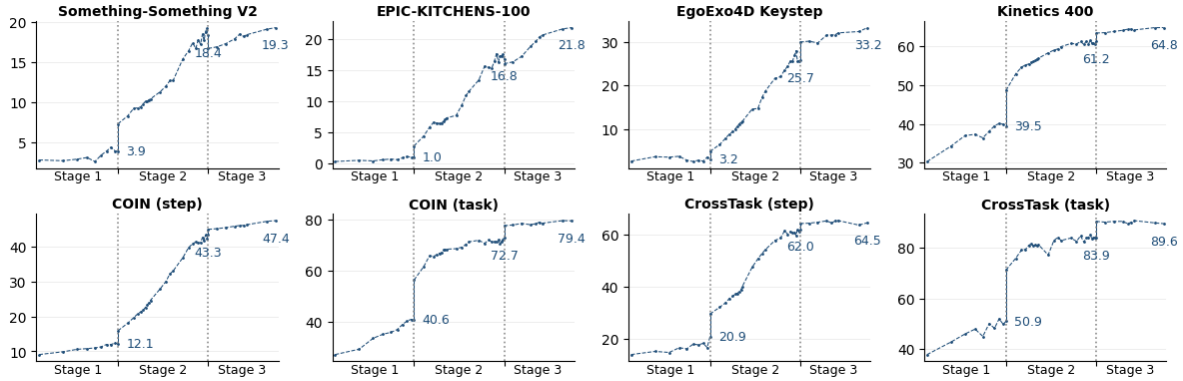


Figure 8. **Zero-shot action recognition accuracy in each training stage.** **Stage 1:** image pretraining with single frame input; **Stage 2:** main Action100M pretraining with 8 frames input; **Stage 2:** 32 frames input with unfrozen encoder. The x-axis of each stage represents number of training samples seen (*i.e.*, number of iterations) in log-scale. The performance at the end of each stage is annotated.

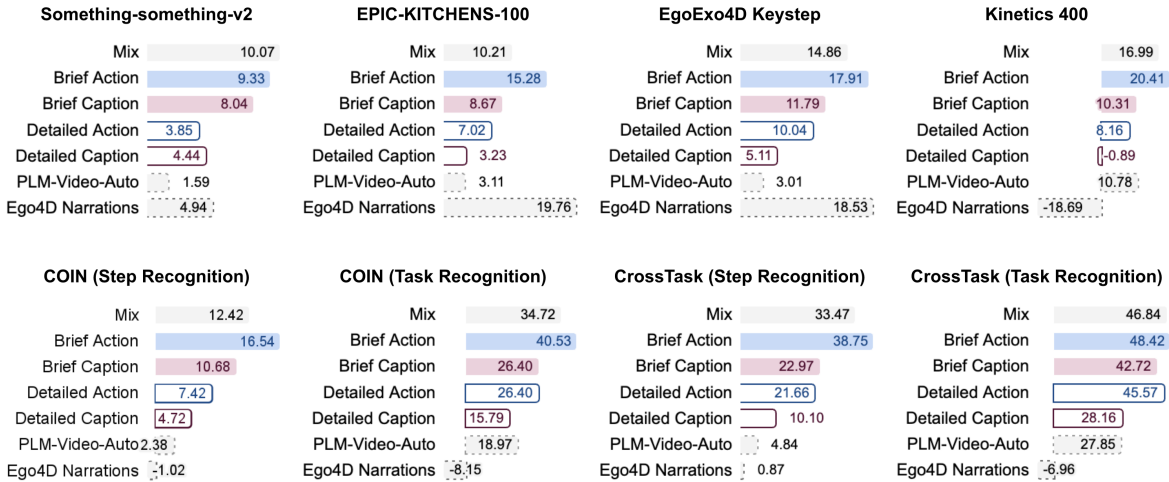


Figure 9. Performance improvements of stage 2 video pretraining with different data upon stage 1.

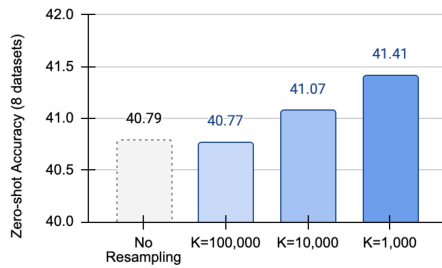


Figure 10. **Effectiveness of semantic resampling.**

report the average across the 8 benchmark zero-shot action classification datasets. As shown in Fig. 10, resampling with a smaller number of clusters leads to improved performance compared to training without resampling, highlighting the benefits of down-sampling frequent actions and up-sampling rare ones.

## 6. Conclusion

We introduced ACTION100M, a large-scale, open-domain dataset for action-centric video understanding built from 1.2M procedural videos and annotated into 147M temporally localized segments. Empirically, pretraining VL-JEPA on ACTION100M yields consistent scaling behavior and strong zero-shot transfer across a diverse set of action recognition and text-to-video retrieval benchmarks, with notable strengths on motion-focused and step-centric datasets. Controlled ablations further show the value of LLM-aggregated brief actions and hierarchical evidence over direct pseudo-labeling, and semantic resampling improves sample efficiency by mitigating long-tail redundancy. Overall, ACTION100M provides a practical route to scale open-vocabulary action understanding, and it enables future work on action anticipation, action-conditioned world models, and long-horizon planning that requires reasoning over multi-scale procedural structure.

## References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1
- [2] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. 1
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1
- [4] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2, 6
- [5] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 1, 2
- [6] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828, 2024. 1
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 1
- [8] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoopfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025. 1
- [9] Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022. 2
- [10] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 6
- [11] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 2, 6
- [12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 2, 6
- [13] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 2, 6
- [14] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2, 6
- [15] Delong Chen, Samuel Cahyawijaya, Etsuko Ishii, Ho Shu Chan, Yejin Bang, and Pascale Fung. What makes for good image captions? *arXiv preprint arXiv:2405.00485*, 2024. 1, 3
- [16] Delong Chen, Willy Chung, Yejin Bang, Ziwei Ji, and Pascale Fung. Worldprediction: A benchmark for high-level world modeling and long-horizon procedural planning. *arXiv preprint arXiv:2506.04363*, 2025. 1
- [17] Delong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722*, 2025. 1, 3
- [18] Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025. 2, 5
- [19] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1

- [20] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. 1, 3
- [21] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 1
- [22] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020. 2, 6
- [23] Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, DeLong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025. 1
- [24] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 6
- [25] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2, 6
- [26] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 6
- [27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 6
- [28] Hilde Kuehne, Ali Arslan, and Thomas Serre. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–787, Los Alamitos, CA, USA, 2014. IEEE Computer Society. 2
- [29] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 1
- [30] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024. 6
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1
- [32] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 1
- [33] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 1, 2, 4, 5
- [34] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 6
- [36] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025. 1
- [37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1
- [38] Fadime Sener, Dibiyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2
- [39] Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*, 2025. 1
- [40] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems*, 36:38863–38886, 2023. 2

- [41] World Labs Technical Staff. Generating worlds. <https://www.worldlabs.ai/blog>, 2024=5. 1
- [42] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 2, 6
- [43] Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *arXiv preprint arXiv:2512.24497*, 2025. 1
- [44] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 6
- [45] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 6
- [46] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025. 7
- [47] Huy V Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024. 7
- [48] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 2, 6
- [49] Qiheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 3
- [50] Wenhao Wang and Yi Yang. VideoUFO: A million-scale user-focused dataset for text-to-video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 2
- [51] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [52] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 1
- [53] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 6
- [54] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 6
- [55] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 2
- [56] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025. 2
- [57] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale. *Advances in Neural Information Processing Systems*, 36: 49428–49444, 2023. 2
- [58] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2, 6
- [59] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 2, 6