VideoEval: Vision-Centric Benchmark Suite for Low-Cost Evaluation of Video Foundation Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

With the accumulation of high-quality data and advancements in visual pretraining paradigms, recent Video Foundation Models (VFMs) have made significant progress, demonstrating remarkable performance on popular video understanding benchmarks. However, conventional benchmarks (e.g. Kinetics) and evaluation protocols are limited by their relatively poor diversity, high evaluation costs, and saturated performance metrics. In this work, we introduce a comprehensive benchmark suite to address these issues, namely **VideoEval**. We establish the **Vid**eo **T**ask **A**daption **B**enchmark (**VidTAB**) and the **Vid**eo **E**mbedding **B**enchmark (**VidEB**) from two perspectives: evaluating the task adaptability of VFMs under few-shot conditions and assessing their feature embedding's direct applicability to downstream tasks. With VideoEval, we conduct a large-scale study of 20 popular open-source vision foundation models. Our study reveals some insightful findings, 1) overall, current VFMs exhibit weak generalization across diverse tasks, 2) increasing video data, whether labeled or in video-text pairs, does not necessarily improve task performance, 3) the effectiveness of some pre-training paradigms may not be fully validated in previous benchmarks, and 4) combining different pretraining paradigms can help develop models with better generalization capabilities. We believe this study serves as a important complement to the current evaluation methods for VFMs and offers valuable insights for future research directions.

1 Introduction

The field of deep learning is experiencing a significant paradigm shift due to the emergence of foundation models (FMs). These models, exemplified by BERT [1], GPT [2, 3, 4], CLIP [5] and Stable Diffusion [6], are trained on massive and diverse data at scale and demonstrate remarkable adaptability to a broad spectrum of downstream tasks.

In the realm of video understanding, early researchers train backbone networks [7, 8, 9, 10] using visual classification tasks on large-scale labeled datasets like ImageNet [11] and Kinetics [12]. However, the high cost associated with labeled data promotes the development of self-supervised learning methods that capitalize on unlabeled data for visual pre-training [13, 14, 15, 16, 17]. Furthermore, researchers delve into multimodal pre-training utilizing large-scale visual-text pairs [18, 19, 20, 21], thereby enhancing their models' capabilities and demonstrating impressive zero-shot performance. Overall, fueled by the accumulation of high-quality image and video data and advancements in visual pre-training paradigms, Video Foundation Models (VFMs) witness remarkable progress in recent years. A new generation of VFMs [15, 16, 22, 23, 24, 25, 26] emerges, demonstrating outstanding performance on conventional video understanding benchmarks.

The rapid development of VFMs raises the problem: ***How to evaluate a video foundation model?*** In image realm, Previous works assess the generalization capability of Image Foundation Models

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

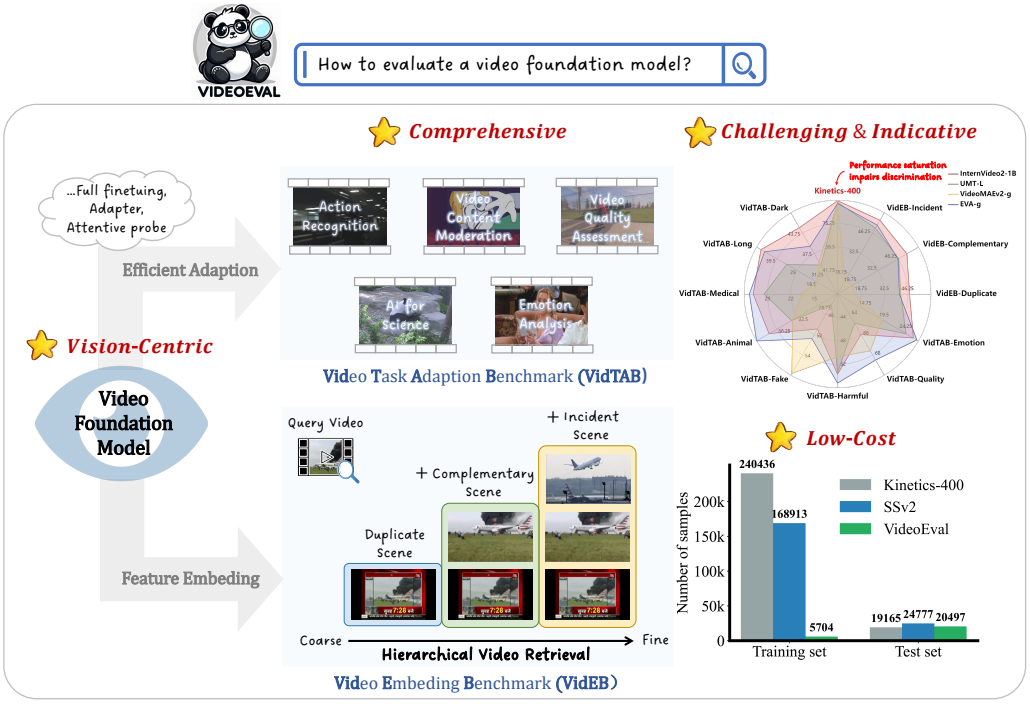

Figure 1: **Overview of VideoEval.** We propose a novel, vision-centric evaluation method for video foundation models that is comprehensive, challenging, indicative, and low-cost.

(IFMs) by evaluating their performance on numerous downstream visual tasks, encompassing diverse scenarios and evaluation protocols [27, 28, 29, 30, 31, 32, 33]. However, previous works primarily evaluates VFMs through benchmarks focusing on action recognition tasks [16, 23, 34]. Some studies [25, 26, 24] have also considered combining language models to evaluate performance on multimodal tasks. There are **several problems with current evaluation methods**: **(1)** Benchmarks like Kinetics [12], Something [35] and AVA [36], which focus on action recognition, overlook other video understanding scenarios (e.g., video quality assessment), limiting their applicability in evaluating the generalization capabilities of visual foundation backbones across diverse video understanding applications. **(2)** The performance of VFMs on conventional benchmark [37] has reached a saturation point (90% Top-1 accuracy), making it challenging to differentiate between the true capabilities of different VFMs. **(3)** The high validation costs associated with conventional evaluation protocols, which often necessitate end-to-end training on the entire dataset, pose a significant challenge, particularly for large VFMs. **(4)** Incorporating language models may introduce bias when evaluating VFMs, as performance differences might stem from the language model rather than the VFMs itself.

To tackle these problems, we build a comprehensive benchmark suite for evaluation of VFMs, namely VideoEval. As shown in Figure 1, our method has the following key features: ***Comprehensive***: First, we created the Video Task Adaptation Benchmark (VidTAB) to evaluate the adaptability of VFMs to unseen tasks with limited samples. We collected public datasets from various video task domains, including action recognition in special scenarios, AI for science, video content moderation, video quality/aesthetic assess, and emotion analysis. From these domains, we constructed eight adaptation tasks and developed evaluation protocols and adaptation methods suitable for current VFMs. Additionally, to assess the capability of VFMs' feature embedding for downstream applications, we created the Video Embedding Benchmark (VidEB), which includes four tasks that evaluate embedding at different granularities. ***Challenging & Indicative***: Due to the diversity of test data and the effectiveness of our evaluation protocols, our VideoEval can effectively distinguish between various VFMs that perform similarly on traditional benchmarks, providing deeper insights into their true capabilities. ***Low-cost***: Thanks to our training-light few-shot evaluation and training-free feature embedding evaluation protocols, VideoEval requires significantly fewer training samples compared to previous benchmarks, while maintaining a comparable number of testing samples to ensure accurate and stable evaluations. ***Vision-centric***: Our evaluation focuses solely on the Video FMs themselves, avoiding the introduction of biases that may arise from incorporating language models.

Based on VideoEval, we evaluate 20 open-source vision foundation models, including VFMs, Image Foundation Models (IFMs), and IFMs with image-to-video methods. **Our main findings as following:** First, current VFMs still struggle to adapt to unseen video tasks with limited training samples. Second, while more data and larger models generally improve performance, augmenting video training data can sometimes negatively affect certain tasks. Third, the effectiveness of certain pre-training paradigms, such as VideoMAEv2 [22], may not have been adequately validated in previous benchmarks. Finally, combining multiple pre-training paradigms can lead to models with better generalization capabilities, such as performing multimodal contrastive learning after unimodal visual self-supervised pre-training [21, 26].

Table 1: **Comparison of VFMs Benchmark.** "Num. training" denotes number of training samples, "Num. test" denotes number of test samples, and "Beyond Action" denotes the tasks in this benchmark extend beyond action understanding. Compared to previous benchmarks, our VideoEval framework achieves more comprehensive and reliable evaluations at a lower cost.

Benchmark	Num. training	Num. test	Beyond Action	Task Diversity	Domain Diversity	VFMs-specific protocol
Single-dataset Benchmarks						
Kinetics-400 [37]	240,436	19,165	✗	✗	✗	✗
Sth-Sth V2 [38]	168,913	24,777	✗	✗	✗	✗
Moment-in-Time [39]	791,246	33,898	✗	✗	✗	✗
UCF101 [40]	9,537	3,783	✗	✗	✗	✗
Multi-dataset Benchmarks						
SEVERE [41]	868,446	144,830	✗	✓	✓	✗
BEAR [42]	240,236	140,436	✗	✓	✓	✗
VideoGLUE [34]	1,896,621	239,011	✗	✓	✓	✓
VideoEval	5,704	20,497	✓	✓	✓	✓

2 Related work

Video foundation models With the continuous growth of image [43, 44, 45] and video data [46, 20, 47, 48, 49] and advancements in pre-training paradigms, research on Video Foundation Models (VFMs) has progressed rapidly. Current VFMs are primarily built around two pre-training paradigms: masked video modeling based on unimodal video data [15, 16, 22, 17, 50, 51, 52] and video-text contrastive learning based on multimodal visual-text pairs [18, 53, 19, 54, 20]. Some works [25, 21, 24] combine these paradigms, enabling VFMs to extend further into multimodal understanding. Additionally, some studies introduce modalities like audio and speech on top of video and text [47, 48, 26], further expanding the capabilities of VFMs. Recently, InternVideo2 [26] leverages mature pre-training paradigms and large-scale high-quality data to scale VFMs to 6 billion parameters, achieving remarkable performance improvements.

Evaluation of VFMs Previous works primarily utilize action recognition benchmarks focused on appearance and motion [12, 38, 36] to evaluate VFMs. To enhance evaluation diversity, some studies explore richer domains and tasks [55, 42, 56], but they remain limited to action recognition tasks. The InternVideo series [25, 26] and VideoGLUE [34] attempt to provide a more comprehensive evaluation of VFMs by expanding the number of benchmarks and evaluation protocols. However, these efforts are still based on existing benchmarks and incurred high validation costs. In contrast, our work considers the characteristics and application scenarios of VFMs, offering a comprehensive and low-cost evaluation solution through task definition and evaluation protocols, aimed at rapidly verifying the generalization capabilities of VFMs—a crucial aspect currently lacking in the community's development of these models.

3 Building VideoEval

We argue that a powerful video foundation model should possess two key capabilities: (1) strong task adaptation ability, i.e., the ability to ***adapt to diverse, unseen tasks with limited training samples***, and (2) the capacity to ***extract feature embedding that retain and distill key information from videos***, directly supporting various downstream tasks. From these perspectives, we construct VideoEval, which includes the Video Task Adaptation Benchmark (VidTAB) and the Video Embedding Benchmark (VidEB). By creating diverse task scenarios and employing efficient evaluation methods, VideoEval can quickly and comprehensively assess the generalization ability of VFMs in video understanding. In

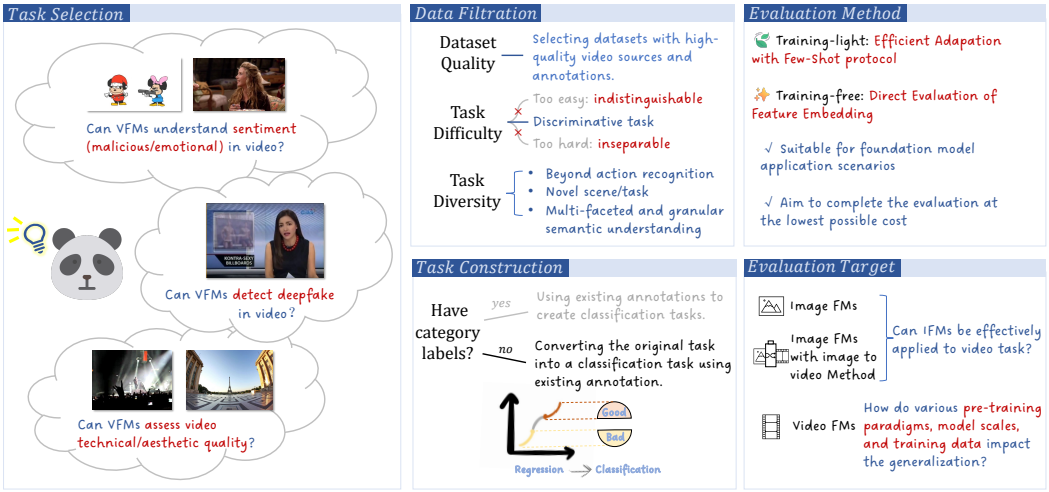

Figure 2: **Illustration of building VideoEval.**

Table 2: **Task details of VideoEval.** All videos are collected from the public datasets for building tasks of VidTAB and VidEB.

Domain	Task	Source	Task Description
Video Task Adaptation Benchmark (VidTAB)			
Action Recognition in Special Scenarios	Action Recognition in Dark Scene	ARID [57]	*Recognizing 11 distinct human actions in dark scenarios.* e.g. Run / Walk / Drink
	Action Recognition in Long Video	BreakFast [58]	*Classifying 10 types of long-duration cooking videos.* e.g. Milk / Tea / Sandwich
AI for Science	Medical Surgery	SurgicalActions160 [59]	*Classifying 16 surgical actions in gynecologic laparoscopy.* e.g. Knotting / Suction / Injection
	Animal Behavior	Animal Kingdom [60]	*Classifying 12 behaviors of wild animals from diverse environmental footage.* e.g. Flying / Chirping / Preening
Video Content Moderation	Fake Face	FaceForensics++ [61]	*Determine whether the faces in the video have been tampered with by AI technology (such as DeepFake).* e.g. Origin video / Video with fake face
	Harmful Content	mob [62]	*Detecting 3 degrees of malicious content within videos.* e.g. Obscene / Indecent activity / Violent activity
Video Quality Assessment	Quality Assess	DOVER [63]	*Evaluating videos from an aesthetic and technical perspective and categorizing them into low and high quality.* e.g. Low quality / High quality
Emotion Analysis	Emotion Analysis	CAER [64]	*Classifying 7 different human emotions in video.* e.g. Happy / Fear / Anger
Video Embedding Benchmark (VidEB)			
Scene Understanding in Temporal Contexts	Duplicate Scene Retrieval	FIVR5K [65]	*Retrieve Duplicate Scene Videos (DSV):* Videos captured by the same camera and sharing at least one scene (without considering any application transformations).
	Complementary Scene Retrieval	FIVR5K [65]	*Retrieve Complementary Scene Videos (CSV):* Retrieve a portion of the same spatiotemporal segment captured from different perspectives.
	Incident Scene Retrieval	FIVR5K [65]	*Retrieving Incident Scene Videos (ISV):* The same event is close in both space and time, but there are no overlapping videos.
	Copy Detection	DVSC23 [66]	*Detecting edited versions of the same source video.* Given a query inserted with one or more copied segments, detect the source video from the database.

this section, we present our VideoEval in detail. The construction pipeline for VideoEval is illustrated in Figure 2, and the evaluation tasks we ultimately constructed are presented in Table 2.

3.1 Video Task Adaption Benchmark

Collecting diverse dataset from public source. Previous benchmarks primarily focus on evaluating video models based on human actions, overlooking many other tasks requiring video understanding. Therefore, we consider five different application scenarios: **1)** *Action Recognition in Special Scenarios* (**Action**): While previous benchmarks have extensively examined action recognition tasks, our focus here is to assess VFMs' capabilities in recognizing actions within special scenarios. **2)** *AI for Science* (**Science**): Referencing previous work [24], we classify tasks related to medicine and natural sciences as a category. **3)** *Video Content Moderation* (**Safety**): We group tasks related to identifying harmful or misleading information in video content. **4)** *Video Quality Assessment* (**Quality**): We categorize more subjective tasks into this group. The goal is to assess VFMs' ability to learn low-level information and human aesthetic preferences. **5)** *Emotion Analysis* (**Emotion**): We group tasks related to human emotion analysis into this category to evaluate VFMs' ability to understand human emotions.

Constructing the adaptation task based on the existing annotations. Classification tasks are straightforward and well-defined, with strong classification performance often indicating robust feature learning. Therefore, they are suitable for evaluating video foundation models. We construct adaptation classification tasks based on the collected data and annotations as follow: **1) Remove Low-Quality Video Datasets**: We manually exclude datasets with videos that have low resolution (below

Table 3: **Task difficulty assessment based on visual language models**. For tasks with fewer categories, such as Fake Face (n=2) and Quality Assess (n=2), random guessing can lead to high accuracy, which may result in a lower apparent proportion of hard samples. Therefore, the zero-shot classification accuracy of the models should also be considered when making task selection.

ratio %	Dark Scene	Long Video	Medical Surgery	Animal Behavior	Fake Face	Harmfull Content	Quality Assess	Emotion Analysis
Easy	18.45	24.57	0.00	19.18	39.06	28.78	53.04	7.21
Spatial	19.00	20.44	4.17	20.86	20.72	24.56	51.24	5.01
Temporal	20.09	22.39	19.79	23.90	4.89	22.76	13.26	27.06
Hard	36.90	26.28	62.50	35.58	9.00	20.17	3.04	47.15

240p), low frame rate (below 15fps), insufficient quantity (fewer than 150 videos per category), or low annotation accuracy (below 90%). **2) Select Discriminative Tasks**: For task difficulty screening, we first evaluate zero-shot classification performance using CLIP-L [5], EVA-g [67], ViCLIP-L [20], and Internvideo2-1B [26]. We then classify samples as follows: *Easy*: Samples that are correctly classified by three or more models. *Spatial*: Samples that are correctly classified by both CLIP and EVA. *Temporal*: Samples that are correctly classified by at least one of ViCLIP or Internvideo2-1B, but not by CLIP and EVA. *Hard*: Samples that are incorrectly classified by all models. We use the zero-shot classification accuracy of the models and the aforementioned proportions as references for task selection. Based on this, we choose tasks with lower zero-shot classification accuracy, higher proportions of Hard and Temporal samples, and lower proportions of Easy samples. The proportions of each type of sample in the tasks we ultimately selected can be found in Table 3. **3) Control the Number of Categories**: For datasets that originally include category labels, such as ARID [57] and Animal Kingdom [60], we select categories with sufficient samples to ensure evaluation accuracy and stability. We also control the final number of categories to avoid making the adaptation task overly difficult. We observed that both zero-shot testing and few-shot experiments based on current VFMs show that when the number of categories is too high, models often perform no better than random guessing. Although this issue may be mitigated as VFMs improve, we currently need to control the number of categories to effectively showcase differences between models. We select the main categories for each task and limit the number of categories to around 10 (based on few-shot experiments). **4) Handling Multi-label and Regression Tasks**: For datasets that are not originally classification tasks, we transform the tasks into classification tasks. For example, for DOVER [63], which is used for video aesthetics and technical quality assessment (a regression task), we assume that videos with quality scores in the top 40% are "high-quality videos" and those with scores in the bottom 40% are "low-quality videos", thus converting the original task into a binary classification task. In total, we construct eight classification tasks to evaluate the adaptation capabilities of video foundation models.

Determining the evaluation protocol. Previous studies [25, 26, 34] typically train video models using entire samples of training set, and most popular benchmarks have large training sample sizes. We argue that this evaluation method overlooks the examination of the adaptation capability of VFMs. As illustrated in Figure 3, under the scenario of using full training samples, the differences between VFMs are difficult to discern. However, under a low-sample protocol, different foundation models exhibit varying degrees of task adaptation capabilities. We observe that for tasks such as Action Recognition in Dark Scenes, which VFMs usually excel at, there are significant differences in adaptation capabilities among different models when training samples are extremely limited (4 shot and 16 shot). As the number of samples gradually increases to 100 shot, these differences diminish. Conversely, for more challenging tasks like Emotion Analysis, the performances of different models are uniformly weak when training samples are extremely limited, showing no discernible differences until a certain number of training samples (100 shot) are reached, at which point different models begin to demonstrate distinct adaptation capabilities. Therefore, to account for the adaptation capabilities of models with different numbers of training samples, we define a task adaptation capability evaluation score (TA-score):

$$TA - score = \frac{Acc^{4s} + Acc^{16s} + Acc^{100s}}{3} \tag{1}$$

Where Acc^{4s}, Acc^{16s}, Acc^{100s} represent the model's top-1 accuracy for 4-shot, 16-shot, and 100-shot classifications, respectively. Unless otherwise specified, we will use TA-score to denote the performance of various tasks in VidTAB.

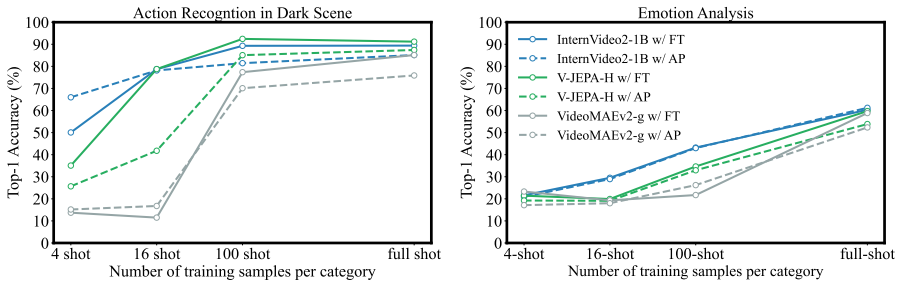

Figure 3: **Performance comparison on different training data scales.** We evaluate the performance variation of multiple video foundation models across tasks from two different domains as the scale of the training data changed. 'FT' and 'AP' denote full finetuning and attentive probe, respectively.

Table 4: **Comparison of adaptation method** All results are obtained using A100-80G with PyTorch-builtin mixed precision, using a batch size of 4 to measure Cuda memory and training time. "Dark" and "Emotion" denote the tasks of Action Recognition in Dark Scenes and Emotion Analysis, respectively. We show the result of V-JEPA-H [23] here,

Adaptation method	Tunable Params (M)	Cuda Memory (G)	Training Time (h)	Dark TA-score	Emotion TA-score
full finetuning	663.7	52.1	1.0	68.8	25.3
adapter	52.6	45.0	1.0	62.4	24.7
attentive probe	19.7	6.4	0.4	54.7	23.8
linear probe	0.0	6.0	0.3	12.9	16.2

Identifying efficient adaptation method for evaluation. We also need to identify how to adapt the foundation models to the corresponding task. Previous work [68, 69, 70, 71, 72] has explored various strategies for efficient adapting the foundation models. Here, we consider several of the most common and popular methods: **Full Finetuning**: Fine-tuning all the parameters of the pre-trained model. **Adapter**: Freezing the pre-trained model and inserting learnable low-rank adapter [73] modules into each block of the pre-trained model for adaptation. **Attentive Probe**: Freezing the pre-trained model and adding an additional learnable cross-attention block at the end of the model to achieve attentive pooling, followed by a linear projection for classification. **Linear Probe**: Directly using the features from the pre-trained model, performing mean pooling, and then using a linear projection for classification. We evaluate the performance of these adaptation methods based on the V-JEPA-H model, as shown in Table 4. Full finetuning and adapter exhibited the best adaptation performance, but incurred high training costs. Linear probe was highly efficient but showed weak adaptation performance. Attentive probe offered a good trade-off between efficiency and adaptation performance. Therefore, in subsequent evaluation experiments, we employed attentive probe to adapt various vision foundation models.

3.2 Video Embedding Benchmark

The main application domains of video embeddings we considering include: Label-Level: Classification and Action Retrieval. Instance-Level: Retrieval, Copy Detection and Ranking. For label-level tasks, VidTAB has already provided a flexible way to evaluate models. Therefore, VidEB aims to assess existing models at a finer semantic level, focusing on instance-level tasks. Although ranking tasks are common in recommendation system scenarios, they are influenced by user information and interactions, in addition to video data. Based on prior research [74], using frozen embeddings for video features does not consistently improve recommendation tasks (resulting in minimal or even negative effects). Thus, we have narrowed the final dataset scope to instance-level retrieval and copy detection. Apart from the traditional classification tasks, the evaluation of representations typically involves standard benchmarks such as video action retrieval [75, 76, 77], which primarily rely on class labels. However, this approach often overlooks the overall scene context and exhibits an overlap with recognition tasks. In contrast, inspired by previous works [78, 66, 79, 80, 81], we establish more rigorous criteria for embedding evaluation in Table 2. Specifically, we require the model to determine the priority and retrieve individual samples based on the overall similarity, rather than solely relying

on class labels. This evaluation protocol provides a more comprehensive assessment of the model's capability to encapsulate subtle visual information.

Evaluation protocol. To facilitate fine-grained embedding evaluation, we incorporate two tasks for assessment: **(1) Hierarchical Video Retrieval** aims to retrieve videos from a database that closely matches the query video in terms of scene, viewpoint, and temporal context. According to previous work [65], videos related to the query are categorized into three levels based on their similarity to the query: Duplicate Scene Videos (DSVs), Complementary Scene Videos (CSVs), and Incident Scene Videos (ISVs), as shown in Table 2: Consequently, the retrieval tasks are structured into three hierarchical levels: *Duplicate Scene Video Retrieval*: only DSVs are positive instances. *Complementary Scene Video Retrieval*: both DSVs and CSVs are positive instances. *Incident Scene Video Retrieval*: DSVs, CSVs, and ISVs are all positive instances. For the evaluation metric, we follow [65] to utilize the mean Average Precision (mAP) to assess the quality of video ranking. **(2) Video Copy Detection** aims to detect edited copies of the query video. Instead of the ranking/retrieval task where all video pairs need to be sorted according to video embedding similarity, it is required to identify a set of video pairs that contain edited versions of the given query. Following [66], we consider the micro-AP (μAP) as our evaluation metric that operates on all queries jointly and takes the confidence scores into account.

4 Benchmarking Video Foundation Models

4.1 Targets and details of evaluation

Evaluation targets We evaluate twenty open-source vision foundation models. Including: (1) twelve video foundation models, covering ***different pre-training paradigms, model scales, and training data scales***, to analyze the impact of these factors on the generalization capability of foundation models. (2) five image foundation models to observe ***how much generalization capability trained on image data can exhibit in video understanding***. (3) three image-to-video methods based on image foundation models to assess the ***effectiveness of current efficient transfer methods***.

Implementation details All models take 8 frames (16 frames if the model has temporal downsampling), with each frame being 224x224 in size as input. For VidTAB, to ensure fair comparison and efficient assessment, we train all models for the same number of epochs and made minor adjustments to the hyperparameters to ensure convergence. For VidEB, all models take 16 frames, with each frame being 224x224 in size as input. In hierarchical video retrieval, the similarity of video-level embedding determines the ranking of retrieval results. In video copy detection, each sample is segmented into 5 clips. The detection confidence score for the entire video is derived from the maximum frame-wise similarity computed for each query-reference pair. See the Appendix for more details.

4.2 Results on VidTAB

Zero-shot evaluation To preliminarily assess the characteristics and difficulty of the dataset, we first evaluate the zero-shot performance of the eight tasks we created using two image language models and two video language models. As shown in the top section of Table 3, both image and video models demonstrated some level of performance for action-related tasks, with video models exhibiting relatively higher performance. For tasks involving low-level information understanding, such as Quality Assessment task, image models performed significantly better. In contrast, for other tasks involving scenarios typically unseen in training data, such as medical surgery videos or Safety Review tasks requiring complex semantic reasoning, all models exhibited almost no performance.

Main results Table 5 presents the evaluation results on VidTAB. We summarize our findings as follows. **On the whole, (1)** Despite exhibiting a degree of generalization capability, *current vision FMs still struggle to adapt to unseen video tasks with limited training samples.* VFMs outperform IFMs, particularly in tasks related to action and behavior understanding. However, IFMs exhibit superior performance on more novel tasks, specifically in the domains of safety and quality, especially when combined with image-to-video adaptation techniques. **(2)** The *adaptation performance of models generally increases with the growth of data and model size*, as observed by the improvements observed from V-JEPA-L to V-JEPA-H (+1.5) and ViCLIP-L-10M to ViCLIP-L-200M (+1.3).

Table 5: **Evaluating state-of-the-art FMs on the VidTAB**. The best and second-best results of foundation models are noted by **blue** and underline, respectively. '**I**', '**V**', and '**IV**' denote image data, video data, and mixed image-video data, respectively. Data marked in gray indicates that the model uses a model trained on that data as initialization. 'K710ft' indicates that the model was fine-tuned with supervision using the labeled action recognition dataset Kinetics-710 (0.66M). Considering the random error in few-shot experiments, we conducted 3-fold experiments for both 4-shot and 16-shot settings, and used their mean as the final result. We also provide the results of full finetuning in the appendix.

	# Params. (M)	# Pt. Data	**Average**	**Action**		**Science**		**Safety**		**Quality**	**Emotion**
				Dark.	Long.	Medical.	Animal.	Harmful.	Fake.	Quality.	Emotion.
Random	-	-	22.7	9.1	10.0	6.3	8.3	33.3	50.0	50.0	14.3
Zero-shot performance of visual language models											
CLIP-L [5]	428	**I**-400M	35.7	29.2	34.6	12.5	32.9	42.1	56.3	65.5	12.9
EVA-CLIP-g [67]	1365	**I**-2B	36.0	32.8	37.2	9.4	28.5	39.6	52.8	69.5	17.9
ViCLIP-L [20]	428	I-400M+**V**-200M	33.6	26.2	37.5	8.3	29.3	32.1	52.2	53.9	29.0
InternVideo2$_{stage2}$ [26]	1350	IV-1.1M+**IV**-25.5M	40.6	37.1	40.2	11.5	45.2	59.1	51.3	56.1	24.3
Image Foundation Model											
CLIP-L [5]	316	**I**-400M	43.2	31.9	37.8	32.3	37.4	54.2	58.2	66.6	27.6
SigLiP-SO [82]	444	**I**-4.11B	43.3	27.6	38.4	36.5	35.8	53.3	58.5	67.8	28.5
EVA-g [83]	1035	**I**-2B	45.8	40.2	47.1	34.4	41.0	51.8	55.2	68.1	29.0
DINOv2-L [84]	317	**I**-142M	42.7	40.8	45.0	39.6	36.1	38.9	52.2	63.2	25.6
DINOv2-g [84]	1165	**I**-142M	44.4	37.8	46.4	42.7	36.0	48.5	53.2	64.3	26.3
*Image Foundation Model with **image-to-video adaptation method***											
ST-Adapter-CLIP-L [70]	328	**I**-400M	46.5	42.4	44.3	31.2	40.1	47.4	64.6	71.5	30.4
AIM-CLIP-L [71]	328	**I**-400M	48.8	41.5	50.0	38.5	40.2	46.4	**69.5**	**73.7**	30.6
ZeroI2V-CLIP-L [72]	303	**I**-400M	46.3	40.3	47.0	31.2	40.2	46.1	65.2	69.9	30.5
Video Foundation Model											
ViCLIP-L-10M [20]	316	I-400M+**V**-10M	41.8	31.2	42.7	30.2	35.3	47.9	53.9	66.2	26.9
ViCLIP-L-200M [20]	316	I-400M+**V**-200M	43.3	38.2	44.6	30.2	37.9	47.4	54.9	65.9	27.5
VideoMAEv1-L [16]	316	**V**-0.24M	43.3	45.6	30.8	31.2	37.4	56.5	51.9	68.7	24.0
VideoMAEv1-H [16]	651	**V**-0.24M	44.7	45.5	31.0	35.4	38.6	55.8	51.8	70.5	29.1
VideoMAEv2-g [22]	1037	**V**-1.35M	37.8	35.2	18.3	18.8	33.7	59.6	50.9	64.7	21.6
VideoMAEv2-g^{k710pt} [22]	1037	**V**-1.35M+K710ft	54.0	**76.4**	72.6	50.0	42.4	43.8	56.9	63.2	27.0
UMT-L$_{stage1}$ [21]	316	**V**-0.66M	40.6	34.3	35.4	30.0	34.2	45.6	53.6	64.7	27.0
UMT-L$_{stage2}$ [21]	316	V-0.66M+**IV**-25M	44.0	34.2	43.9	22.9	39.4	**63.9**	53.0	67.3	27.4
V-JEPA-L [23]	318	**V**-2M	43.5	50.4	34.3	39.6	39.7	43.9	51.7	66.7	21.4
V-JEPA-H [23]	653	**V**-2M	45.1	53.8	37.6	35.4	40.4	47.3	53.0	68.1	25.1
InternVideo2-1B$_{stage1}$ [26]	1037	**IV**-1.1M	46.1	45.2	50.3	33.3	38.7	52.3	53.5	65.9	29.3
InternVideo2-1B$_{stage1}$ [26]	1037	**IV**-1.1M+K710ft	**56.7**	75.6	**77.5**	**53.1**	45.4	47.2	55.5	66.2	**33.2**
InternVideo2-1B$_{stage2}$ [26]	1037	IV-1.1M+**IV**-25.5M	53.6	66.0	71.1	38.5	**50.0**	53.6	54.7	64.3	30.3

For the pre-training data, (3) *While augmenting video training data is generally beneficial, it can negatively impact the performance on some tasks.* For both VideoMAEv2-g and InternVideo2-1B$_{stage1}$, fine-tuning on Kinetics-710 data significantly enhances Action-related tasks, but consistently degrades certain Safety and Quality tasks. Similar findings are observed with ViCLIP-L, where post-pretraining on a large-scale video dataset improves Action-related tasks but diminishes performance in other domains (Science, Safety, Quality, Emotion). It could be attributed to the limited diversity of the current video training data. **(4)** For models trained on single-modal visual data, *incorporating additional weak-supervised post-pretraining with visual-text data leads to significant improvements in adaptation capabilities*. This is evident in the performance gains observed from UMT-L$_{stage1}$ to UMT-L$_{stage2}$ (+3.6) and from InternVideo2-1B$_{stage1}$ to InternVideo2-1B$_{stage2}$ (+8.0). Interestingly, this finding contradicts previous conclusions drawn from commonly used action recognition benchmarks, suggesting that these benchmarks may introduce bias. **For the pre-training paradigms of model, (5)** *The effectiveness of pre-training paradigms in scaling model size might not be adequately validated on popular action recognition benchmarks.* While VideoMAEv2 successfully scaled a model to 1B parameters using the dual masking strategy [22], its adaptation performance (37.7 vs 44.4) significantly declined compared to VideoMAEv1-H. Interestingly, VideoMAEv2-g demonstrated remarkable performance after fine-tuning on Kinetics-710 (0.66M), suggesting that the abundant labeled data may have compensated for the shortcomings of its pre-training performance. **(6)** *Single-modal self-supervised pre-training paradigms exhibit superior data efficiency compared to multimodal weakly-supervised pre-training paradigms*. Notably, V-JEPA and VideoMAEv1, trained solely on relatively small-scale unlabeled video data via self-supervised pre-training, demonstrate comparable or even superior performance to ViCLIP, which is trained on a massive dataset of video-

Table 6: **Evaluation of State-of-the-Art Foundation Models on the VidEB Dataset.** "K400pt" and "K400ft" denote that the model is pre-trained and fine-tuned, respectively, using the labeled action recognition dataset Kinetics-400 (0.31M). MCL: Multi-modal Contrastive Learning, SCL: Self-supervised Contrastive Learning, MVM: Masked Video Modeling, SFT: Supervised Fine-tuning. Other notations are consistent with those in Table 5.

	Pretrain Tasks	# Pretrain Data	**Average**	**Scene**			
				Duplicate	Complementary	Incident	Copyright
Image Foundation Model							
CLIP-L [5]	MCL	**I**-400M	43.0	41.1	46.4	52.0	32.3
EVA-g [83]	MCL	**I**-2B	37.1	41.4	46.1	51.7	9.3
SigLiP-SO [82]	MCL	**I**-4.11B	38.6	40.6	45.5	51.5	16.9
DINOv2-L [84]	SCL	**I**-142M	45.6	49.0	53.5	54.3	25.6
DINOv2-g [84]	SCL	**I**-142M	48.6	**50.5**	**55.1**	**56.0**	32.8
Video Foundation Model							
VideoMAEv1-L [16]	MVM	K400pt	12.9	14.5	15.1	13.2	8.8
VideoMAEv1-L-K400ft [16]	MVM+SFT	K400pt+ft	27.4	27.6	30.2	30.3	21.6
VideoMAEv2-g [22]	MVM	**V**-1.35M	11.6	14.8	15.4	13.4	2.8
VideoMAEv2-g-K710ft [22]	MVM+SFT	**V**-1.35M+K710ft	37.4	33.8	37.1	37.1	41.7
UMT-L$_{stage1}$ [21]	MVM	**V**-0.66M	41.1	42.2	46.6	49.6	25.7
UMT-L$_{stage1}$-K710ft [21]	MVM+SFT	**V**-0.66M+K710ft	29.0	26.4	29.4	30.3	30.0
UMT-L$_{stage2}$ [21]	MVM+MCL	V-0.66M+**IV**-25M	34.2	33.4	37.3	40.6	25.4
V-JEPA-L [23]	MVM	**V**-2M	19.7	21.3	23.9	21.7	12.0
V-JEPA-H [23]	MVM	**V**-2M	20.2	21.5	23.7	21.2	14.3
InternVideo2-1B$_{stage1}$ [26]	MVM	**IV**-1.1M	**50.4**	47.3	52.1	54.9	**47.3**
InternVideo2-1B$_{stage1}$-K710ft [26]	MVM+SFT	**IV**-1.1M+K710ft	33.9	30.5	34.2	34.1	36.9
InternVideo2-1B$_{stage2}$ [26]	MVM+MCL	IV-1.1M+**IV**-25.5M	34.6	32.4	36.8	39.9	29.3

text pairs. **In addition, (7)** *Effective adaptation method for FMs is crucial*. Three image-to-video methods based on CLIP-L achieved significant performance improvements compared to using an attentive probe directly. We believe this represents a promising avenue for future research.

4.3 Results on VidEB

The main results of VidEB are presented in Table 6. We evaluate the embedding performance using different pre-training paradigms for IFMs and VFMs as frozen feature extractors. Surprisingly, **IFMs performs better than most VFMs**, likely due to the existing gap in spatial modeling capabilities between VFMs and IFMs. **For the pre-training paradigms of the model, (1)** *The contrastive learning (CL) based approach consistently excels in embedding evaluation.* Due to CL's emphasis on the relationships between samples during training, DINOv2, which focuses solely on vision, outperforms vision-language contrastive methods like CLIP across multiple tasks. **(2)** *The effectiveness of masked video modeling is closely tied to the targets it reconstructs or aligns with.* With higher semantic richness, it shows progressive improvements in embedding quality for VideoMAE-L, V-JEPA-L, and UMT-L$_{stage1}$. **(3)** *Vision-centric pretraining outperforms Multi-modal pretraining in vision-centric scenarios.* Comparing UMT-L$_{stage1}$ and InternVideo2-1B$_{stage1}$ with their multi-modal counterparts UMT-L$_{stage2}$ and InternVideo2-1B$_{stage2}$, the introduction of visual-text pair data in multi-stage training does not enhance performance in vision-centric scenarios. This is also consistent with the performance differences observed between DINO and CLIP-style pre-training methods. Additionally, we assess the **impact of fine-tuning on the embedding evaluation of these pre-trained models**. **(4)** *Labels bring new semantic information or disrupt existing finer-grained semantic information.* The performance variations after fine-tuning differ based on the pre-training strategy. For UMT-L$_{stage1}$ and InternVideo2-1B$_{stage1}$, fine-tuning leads to a significant drop in performance (-12.1 for UMT and -16.5 for InternVideo) due to the introduction of more singular label information, which causes catastrophic forgetting. In contrast, VideoMAE and VideoMAEv2 show substantial performance gains (+14.5 and +25.8, respectively) because the low-level semantics learned during pre-training are less abstract and benefit more from the addition of high-level label information.

5 Conclusions

We present VideoEval, a comprehensive benchmark suite for efficiently evaluating the VFMs. To this end, we establish VidTAB, which explores suitable evaluation tasks and protocols for VFMs from the perspective of assessing their adaptability to unknown tasks with limited samples. Additionally, we create VidEB to evaluate the capability of VFMs' feature embedding in directly supporting downstream tasks. Utilizing VideoEval, we conduct a large-scale study involving 20 popular open-source vision foundation models, providing valuable insights for future research directions.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2018.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[3] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[4] OpenAI. Gpt-4v(ision) system card. `https://api.semanticscholar.org/CorpusID:263218031`, 2023.

[5] 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 *ICML*, 2021.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[9] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022.

[10] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[12] 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*, abs/1705.06950, 2017.

[13] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, 2021.

[14] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.

[15] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *NeurIPS*, 2022.

[16] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

[17] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. *CVPR*, 2022.

[18] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021.

[19] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Video-text modeling with zero-shot transfer from contrastive captioners. *ArXiv*, abs/2212.04979, 2022.

[20] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Jian Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Y. Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *ICLR*, 2024.

[21] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, 2023.

[22] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023.

[23] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

[24] Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. *ArXiv*, abs/2402.13217, 2024.

[25] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *ArXiv*, abs/2212.03191, 2022.

[26] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *ArXiv*, abs/2403.15377, 2024.

[27] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. The visual task adaptation benchmark. *Arxiv*, abs/1910.04867, 2019.

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[29] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[30] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

[31] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[32] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. In *ICLR*, 2023.

[33] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. In *NeurIPS*, 2023.

[34] Liangzhe Yuan, Nitesh Bharadwaj Gundavarapu, Long Zhao, Hao Zhou, Yin Cui, Lu Jiang, Xuan Yang, Menglin Jia, Tobias Weyand, Luke Friedman, et al. Videoglue: Video general understanding evaluation of foundation models. *ArXiv*, abs/2307.03166, 2023.

[35] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[36] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. *CVPR*, 2017.

[37] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017.

[38] 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 *ICCV*, 2017.

[39] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *TPAMI*, 2020.

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *Arxiv*, abs/1212.0402, 2012.

[41] Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, and Cees GM Snoek. How severe is benchmark-sensitivity in video self-supervised learning? In *ECCV*, 2022.

[42] Andong Deng, Taojiannan Yang, and Chen Chen. A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. In *ICCV*, 2023.

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[44] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[46] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[47] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. VALOR: vision-audio-language omni-perception pretraining model and dataset. *Arxiv*, abs/2304.08345, 2023.

[48] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *NeurIPS*, 2024.

[49] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *Arxiv*, abs/2402.19479, 2024.

[50] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, 2023.

[51] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *CVPR*, 2023.

[52] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023.

[53] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. In *CVPR*, 2023.

[54] Feng Cheng, Xizi Wang, Jie Lei, David J. Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. *ArXiv*, abs/2212.05051, 2022.

[55] Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, and Cees G. M. Snoek. How severe is benchmark-sensitivity in video self-supervised learning? In *ECCV*, 2022.

[56] Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S Rawat. A large-scale robustness analysis of video action recognition models. In *CVPR*, 2023.

[57] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *IJCAI*, 2021.

[58] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014.

[59] Klaus Schoeffmann, Heinrich Husslein, Sabrina Kletz, Stefan Petscharnig, Bernd Muenzer, and Christian Beecks. Video retrieval in laparoscopic video recordings with dynamic content descriptors. *Multimedia Tools and Applications*, 77:16813–16832, 2018.

[60] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *CVPR*, 2022.

[61] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

[62] Syed Hammad Ahmed, Muhammad Junaid Khan, HM Qaisar, and Gita Sukthankar. Malicious or benign? towards effective content moderation for children's videos. *ArXiv*, abs/2305.15551, 2023.

[63] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023.

[64] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, 2019.

[65] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Fivr: Fine-grained incident video retrieval. *IEEE Transactions on Multimedia*, 21, 2019.

[66] Ed Pizzi, Giorgos Kordopatis-Zilos, Hiral Patel, Gheorghe Postelnicu, Sugosh Nagavara Ravindra, Akshay Gupta, Symeon Papadopoulos, Giorgos Tolias, and Matthijs Douze. The 2023 video similarity dataset and challenge. *Computer Vision and Image Understanding*, 2024.

[67] Quan Sun, Yuxin Fang, Ledell Yu Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *ArXiv*, abs/2303.15389, 2023.

[68] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

[69] Bruce XB Yu, Jianlong Chang, Haixin Wang, Lingbo Liu, Shijie Wang, Zhiyu Wang, Junfan Lin, Lingxi Xie, Haojie Li, Zhouchen Lin, et al. Visual tuning. *ACM Computing Surveys*, 2023.

[70] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. Parameter-efficient image-to-video transfer learning. *arXiv*, abs/2206.13559, 2022.

[71] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. In *ICLR*, 2023.

[72] Xinhao Li and Limin Wang. Zeroi2v: Zero-cost adaptation of pre-trained transformers from image to video. *ArXiv*, abs/2310.01324, 2023.

[73] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *EMNLP*, 2020.

[74] Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. A content-driven micro-video recommendation dataset at scale. *arXiv preprint arXiv:2309.15379*, 2023.

[75] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *NeurIPS*, 2020.

[76] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019.

[77] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020.

[78] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[79] Xiao Wu, Alexander G Hauptmann, and Chong-Wah Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 218–227, 2007.

[80] Yu-Gang Jiang, Yudong Jiang, and Jiajun Wang. Vcdb: a large-scale database for partial copy detection in videos. In *ECCV*, 2014.

[81] Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoë Papakipos, Lowik Chanussot, Filip Radenovic, Tomas Jenicek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, et al. The 2021 image similarity dataset and challenge. *arXiv preprint arXiv:2106.09672*, 2021.

[82] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[83] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023.

[84] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We place it in Appendix Section D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper don't have theory assumptions and proofs

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See `https://github.com/MCG-NJU/VideoEval`.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See `https://github.com/MCG-NJU/VideoEval`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Appendix and `https://github.com/MCG-NJU/VideoEval`.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For VidTAB, we referred to the setting of the previous few shot work and repeated the few shot experiment three times to reduce randomness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See `https://github.com/MCG-NJU/VideoEval` for training details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper follows the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix Section D.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We don't provided new model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: See `https://github.com/MCG-NJU/VideoEval`, we provide our annotations here.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only use LLM for simple work like writing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.