Foundation Models for Video Understanding: A Survey

Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S. Rawat, and Thomas B. Moeslund,

给based model 分类

Abstract—Video Foundation Models (ViFMs) aim to develop general-purpose representations for various video understanding tasks by leveraging large-scale datasets and powerful models to capture robust and generic features from video data. This survey analyzes over 200 methods, offering a comprehensive overview of benchmarks and evaluation metrics across 15 distinct video tasks, categorized into three main groups. Additionally, we provide an in-depth performance analysis of these models for the six most common video tasks. We identify three main approaches to constructing ViFMs: 1) Image-based ViFMs, which adapt image foundation models for video tasks; 2) Video-based ViFMs, which utilize video-specific encoding methods; and 3) Universal Foundation Models (UFMs), which integrate multiple modalities (image, video, audio, text, etc.) within a single framework. Each approach is further subdivided based on either practical implementation perspectives or pretraining objective types. By comparing the performance of various ViFMs on common video tasks, we offer valuable insights into their strengths and weaknesses, guiding future advancements in video understanding. Our analysis reveals that image-based ViFMs consistently outperform video-based ViFMs on most video understanding tasks. Additionally, UFMs, which leverage diverse modalities, demonstrate superior performance across all video tasks. We provide the comprehensive list of ViFMs studied in this work at: https://github.com/NeeluMadan/ViFM_Survey.git.

Index Terms—Foundation models, Self-supervised learning, Large-scale Pretraining, Large Language Models, Video Understanding

1 Introduction

The increasing availability of powerful computing resources and ever-growing datasets has fueled the development of foundation models [1], [2]. These versatile AI models, trained on massive amounts of data using self-supervised or semi-supervised learning, can be fine-tuned for various downstream tasks. Initial successes focused on static images [3], [4], with models like CLIP [3], and SAM [5] achieving impressive results. Recent research [6], [7] has extended this to video domain, where several pretraining strategies have been developed for Video Foundational Models (ViFMs).

While video analysis and generation has been of interest to the computer vision community for decades [8], [9], [10], [11], [12], [13], and the problem has largely been challenging due to complexity in tasks, additional time dimension, and volume of data. The initially developed approaches are mostly based on processing individual frames with standard image analysis techniques and additional temporal aspect on top [9], [14]. Alternatively, more advanced techniques were developed specifically designed for videos, such as 3D convolutions [15], recurrent networks [15], use of optical-flow [16], and transformers [11], [17], operating directly on videos providing better temporal modeling Furthermore, there has been significant research exploring the role of multiple modalities to enhance video understanding [18], [19].

We see a similar trend in ViFMs and their evolution also follows extending images (Image-based ViFMs), separate video modeling (Video-based ViFMs), and incorporating additional modalities, e.g., Automatic Speech Recognition (ASR) (Universal FMs).

Motivation and Contribution. The field of video understanding is undergoing significant advancement, as evident by the increasing number of research publications focused on various video understanding tasks (Figure 1). This growth coincides with the development of large-scale pretraining techniques. These techniques have demonstrated remarkable capabilities in adapting to diverse tasks, requiring minimal additional training with robust generalization. As a result, researchers are actively investigating the role of these foundational models to address a broad spectrum of video understanding challenges. To navigate this rapidly evolving research landscape (See Figure 3), a systematic review of video understanding models is essential. We attempt to fill this critical gap by providing a comprehensive analysis of foundational models employed in video understanding tasks. We hope that this survey helps to provide a roadmap for future research directions associated with video understanding.

Related Surveys. While several surveys have delved into specific video understanding tasks [20], [21] or foundational models for images [2], with surveys such as Shiappa et al. [22], which offers an extensive review of self-supervised approaches for video understanding, the landscape has evolved significantly in recent years. With the rise of large-scale foundational models, there is a need for a comprehensive review specifically focused on these models in the context of video understanding. To the best of our knowledge, our

- *N. Madan is with the Visual Analysis and Perception Lab, Aalborg University, Denmark,*
- *A. Møgelmose, and T. B. Moeslund is with the Visual Analysis and Perception Lab, Aalborg University, Denmark, and Pioneer Center for AI,*
- *R. Modi, and Y. S. Rawat is with the Center for Research in Computer Vision (CRCV), Department of Computer Science, University of Central Florida, Orlando, FL, 32816.*

Corresponding author: Neelu Madan (nema@create.aau.dk)

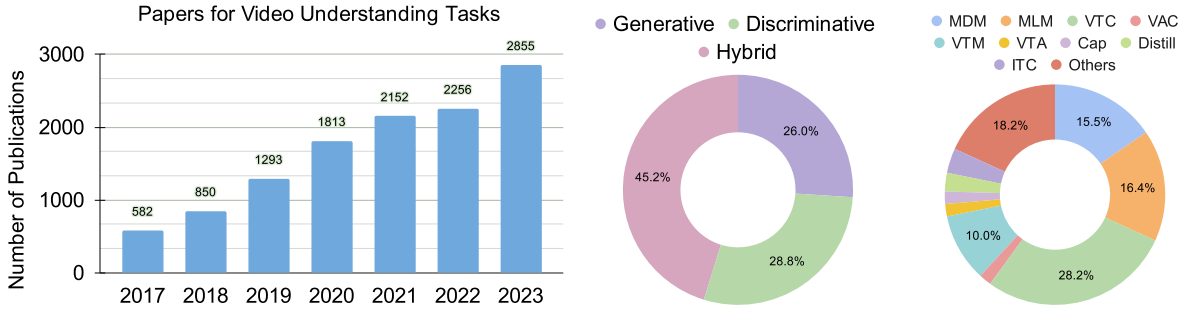


Fig. 1: Overview of recent research trends in video understanding. The **left bar chart** shows a significant increase in publications on this topic, based on data from prestigious conferences and journals. The figure presents statistics showcasing research focusing on generative, discriminative, and hybrid pretraining objectives, as depicted in the **center pie chart**. Specific pretraining objectives such as Mask Data Modeling (MDM), Mask Language Modeling (MLM), Vision-Text Contrastive (VTC), Vision-Audio Contrastive (VAC), Vision-Text Matching (VTM), Vision-Text Alignment (VTA), Captioning Loss (CAP), and Distillation Loss (Distill) are highlighted in the **right pie chart**. Best viewed in color.

survey is the first to provide such a comprehensive overview of foundation models for video understanding. The main contributions of our survey are:

- This work presents the first comprehensive survey of foundational models (ViFMs) deployed for diverse video understanding tasks. Our survey categorizes ViFMs into three groups: i) *Image-based ViFMs:* Trained solely on image data. ii) *Video-based ViFMs:* Leveraging video data during training. iii) *Universal Foundation Models (UFMs):* Combining various modalities (image, video, audio, text) during pretraining.
- We uniquely categorize video understanding tasks by their primary focus and temporal involvement, providing an extensive list of datasets and evaluation metrics for each task.
- We conduct a comprehensive comparison of ViFMs from each category, analyzing various research findings. This analysis reveals valuable insights regarding the most effective ViFMs for different video understanding tasks.
- This survey further identifies crucial challenges faced by ViFMs, highlighting open problems that require further exploration. Additionally, we discuss promising future directions for ViFM development, paving the way for advancements in video understanding.

Paper Organization. In the first part of the paper (section 2), we cover a wide range of video analysis tasks, ranging from video classification to generation. We discuss widely used architectures and loss functions, as well as datasets relevant for large-scale pretraining. Next, we explain the main categories of ViFMs namely: Image-based ViFMs (Sec 3), Video-based ViFMs (Sec 4), and Universal FMs (Sec 5) (See Figure 2 for the taxonomy). Finally (sections 6-7), we compare and discuss the performance of the presented models, as well as present challenges and future directions for the field.

2 Preliminaries

In this section, we lay the groundwork for understanding this survey. We begin by defining the diverse tasks involved in video understanding, allowing the reader to grasp various goals and challenges associated with analyzing video data. Next, we delve into the major architectural styles adopted by different foundation models. Finally, we offer a concise overview of the large-scale pretraining process, the necessary datasets used for training, and the methods used to adapt these generic models for different video tasks.

2.1 Video Understanding Tasks

In this section, we discuss video tasks (see Fig. 5) focused on video content understanding, descriptive understanding, and content generation, along with their popular benchmarks and evaluation metrics.

2.1.1 *Video Content Understanding*

Computer vision tasks for video content understanding fall into three primary levels. *a) Abstract understanding* focuses on inferring the video's overall theme or activity (e.g., video retrieval, action recognition). *b) Temporal understanding* involves determining specific time-related aspects of the video (e.g., temporal action localization). *c) Spatio-temporal understanding* In addition to temporally analyzing the order of events happening in the video, this level of understanding also considers their precise location in each frame. By traversing this progression, deep models build a sophisticated comprehension of video content, similar to how humans gradually grasp information.

a) Abstract Understanding Tasks.

Action Recognition. The task is to assign a category to a video. Different benchmarks for action recognition are: Kinetic-400 [23], Kinetic-600 [24], Kinetic-700 [25], Something-Something-V1 (SSv1) [26], Something-Something-V2 (SSv2) [26], ActivityNet [27], HACS [28], HMDB51 [29], UCF-101 [30], TinyVIRAT [31], and Diving-48 [32]. This task is evaluated using Top-K accuracy as a metric.

Retrieval. The task involves finding videos containing specific actions, objects, or scenes. This task exists in literature as: *i) Text-to-video (T2V) retrieval:* T2V retrieval [33] finds videos using textual descriptions (single sentence). Common benchmarks for this task are EPIC-Kitchen-100 [34], ActivityNet Captions [35], QuerYD [36], CondensedMovie [37], Kinetic-Geb [38], MSRVTT [39], DiDeMo [40], YouCook2 [41], and LSMDC [19]. The primary evaluation metric for this task is Recall at K (R@K), with R@1 indicating the accuracy of the top retrieved result [33]. *ii) Multi-Instance Retrieval (MIR):* MIR

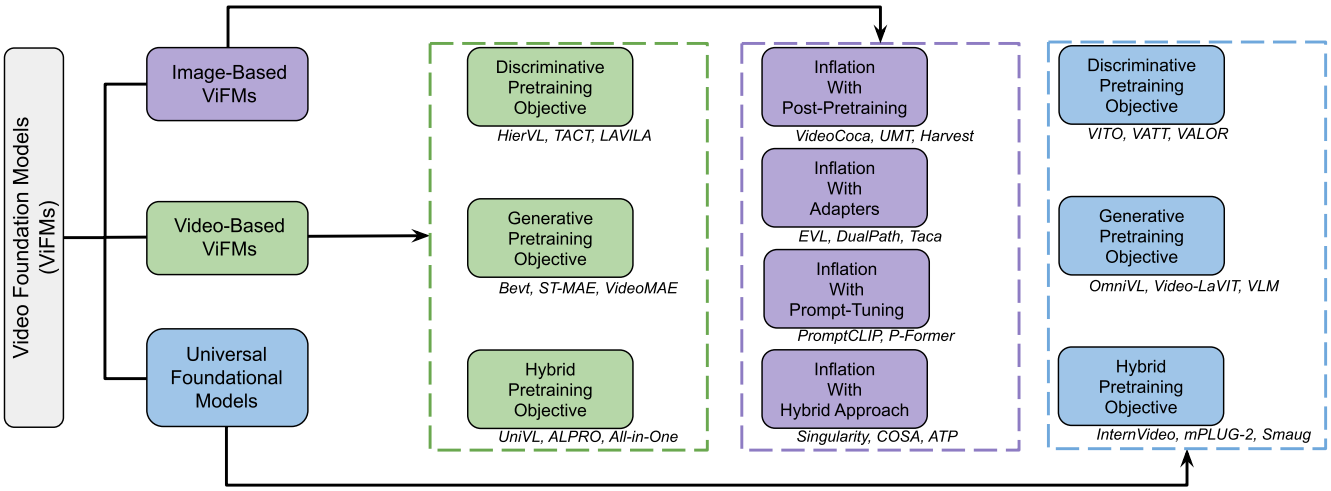


Fig. 2: Figure illustrates the classification of ViFMs into three categories: **a) Image-based ViFMs**, encompassing Inflation with Post-pretraining (3.1), Adapters (3.2), Prompt-tuning (3.3), and Hybrid Approaches (3.4); **b) Video-based ViFMs**, including Generative (4.1), Discriminative (4.2), and Hybrid (4.3) pretraining objectives; **c) Universal Foundational Models (UFMs)**, comprising Generative (5.1), Discriminative (5.2), and Hybrid (5.3) pretraining objectives. Best viewed in color.

[42] combines both text-to-video (T2V) retrieval and video-to-text (V2T) retrieval, where the goal is to find a video given its textual description or vice versa. A common benchmark for this task is EPIC-Kitchen-100 [43]. The evaluation metrics for MIR include mean Average Precision (mAP) and normalized Discounted Cumulative Gain (nDCG) [42].

Video Anomaly Detection (VAD). VAD aims to identify anomalies in video sequences. Common benchmarks for this task include Avenue [44], ShanghaiTech [45], Street-Scene [46], UBNormal [47], and UCF-Crime [48] datasets. Evaluation metrics include Area-under-the-Curve (AUC), Region-based Detection Criteria (RBDC), and Track-based Detection Criteria (TBDC).

b) Temporal Understanding Tasks.

Fine-grained Classification. This task extends the classification task for long-form videos [49], where COIN [50] and LVU [51] are the benchmarking datasets. COIN [50] proposes Procedural Activities Classification (PAC), where the task is to divide complex actions into meaningful subactions and then learn the correct order and hierarchical-relationship among these subactions. LVU [51] proposes 9 tasks including content understanding (relationship, speaking style, scene/place) for the comprehensive video understanding. These fine-grained classification tasks thus require spatio-temporal understanding of videos.

Temporal Action Localization (TAL). TAL [52] aims to pinpoint the exact moments within videos where specific actions occur. Common benchmarks for these tasks are THUMOS-14 [53], ActivityNet-v1.3 [27], HACS Segment [28], FineAction [54], BreakFast [55], Charades [56], and Ikea-ASM [57]. The evaluation metric is the mean average precision (mAP) and the average precision (AP) [27], [53] for each action category.

c) Spatio-temporal Understanding Tasks.

Spatio-temporal Action Localization (STAL). STAL aims to find both ”when” and ”where” specific actions unfold within a video [52]. Notable datasets for this particular category are UCF101-24 [30], JHMDB-22 [58], and UCF-MAMA [59]. These datasets contain annotation for each video frame. Datasets like Ava [60], and Ava-Kinetics [61] contain box-annotations at 1Hz sampling frequency over a clip of 15 mins. Evaluation metric for this task is the f-mAP and video-mAP [62] measuring frame-level and video-level localization performance respectively. (mAP: mean average precision)

Tracking. The task aims at identifying and following the movement of objects throughout a video. KITTI [63], UA-DETRAC [64], LaSOT [65], MOT16/MOT17 [66], MOT20 [67], MOTSynth [68], BDD-100K [69], TAO [70], BURST [71], and LV-VIS [72] are popular benchmarking datasets; and HOTA [73], and Clear-MOT [74] are evaluation metrics for this task. Recently, a more fine-grained tracking approach known as *point tracking* has emerged, which tracks specific points on an object’s surface regardless of pixel location. Datasets for point tracking include PointOdyssey [75], TAP-Vid-DAVIS [76] and CroHD [77], while evaluation focuses on Mean Trajectory Error (MTE) and position accuracy.

Video Object Segmentation (VOS). VOS aims to segment and track objects in video sequences. Benchmarks for different segmentation types include Youtube-VOS for video object segmentation [78], Youtube-VIS and DAVIS for video instance segmentation [79], [80], [81], and CityScapes-VPS [82] for video panoptic segmentation.

A variant of VOS called Referring Video Segmentation (RVS) combines vision and text modalities. RVS segments objects referred to by textual descriptions or the first frame’s segmentation. Benchmarks for RVS using textual descriptions include RefCOCOg [83], Refer-Youtube-VOS [84], Refer-DAVIS [85], A2D-Sentences [86], [87], and JHMDB-Sentences [86]. Benchmarks for RVS using the first frame’s segmentation as a reference include Youtube-VOS and DAVIS [78], [81].

Common evaluation metrics for video segmentation tasks are mean Average Precision (meanAP) and mean Intersection over Union (meanIoU), except panoptic segmentation quality (PSQ) [82] for video panoptic segmentation, and region (J) and boundary (F) metrics [81] for video instance segmentation.

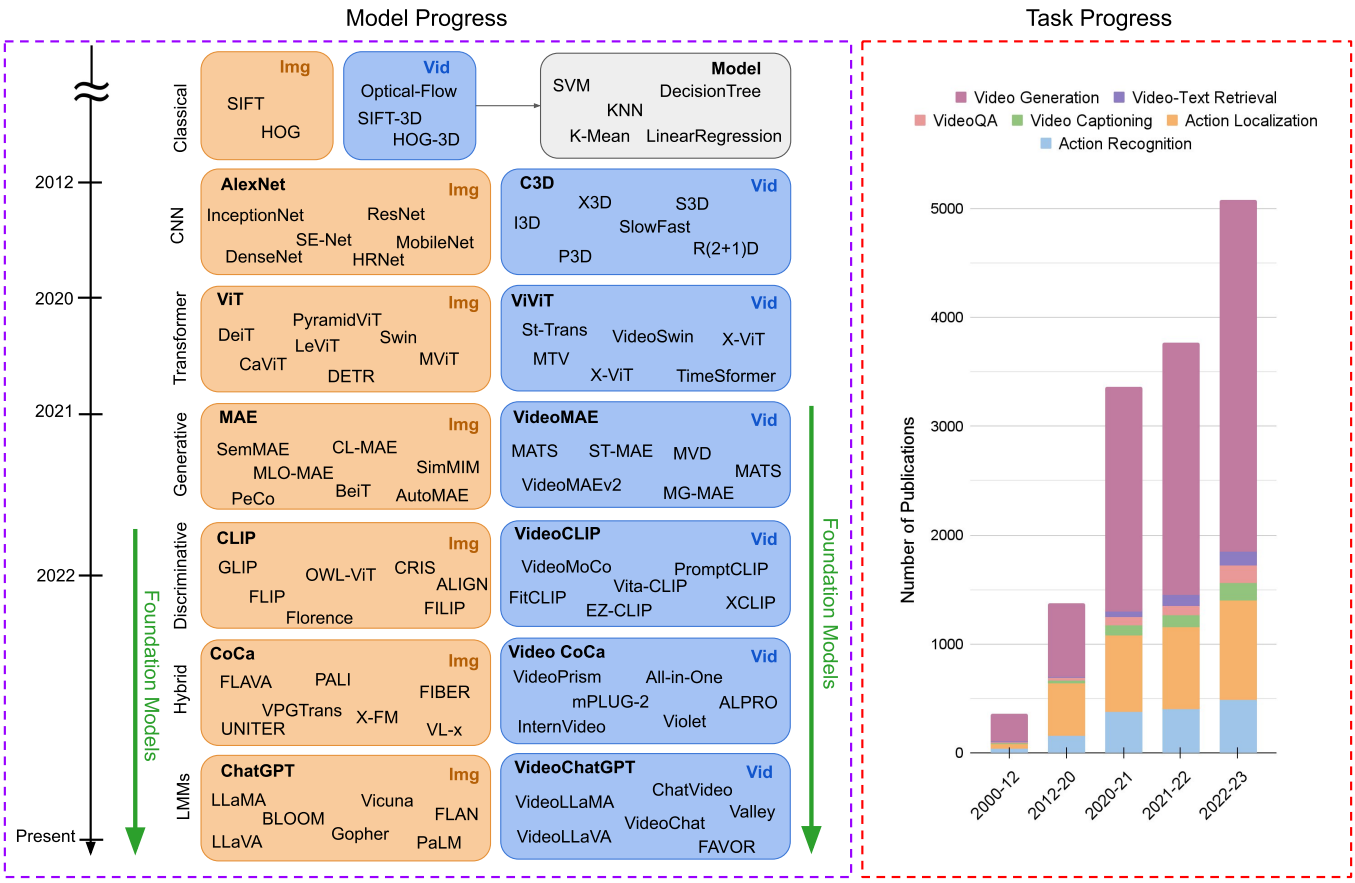


Fig. 3: Figure contrasts classical (separate feature extraction, model training) and deep learning (unified framework) approaches in computer vision. It also shows the progression of deep learning approaches for both image and video processing over time. Best viewed in color.

Video Object Detection (VOD). VOD aims to detect objects across a video stream. Benchmarks for this task include ImageNet-VID [88] and EPIC-Kitchen [34]. A multi-modal variant of this task, called "Spatio-temporal Video Grounding" (STVG), localizes objects referred to by textual descriptions. Benchmarks for STVG include VidSTVG [89] and HC-STVG [90]. Evaluation metrics for VOD tasks are mean Average Precision (mAP) and mean Intersection over Union (meanIoU).

2.1.2 Descriptive Understanding Tasks

This section covers benchmarks and evaluation metrics associated with Video Question Answering (VideoQA), and Video Captioning tasks. Both VideoQA and Captioning tasks focus on understanding of the textual description of the video content.

Video question answering (VideoQA). VideoQA answers questions about the video content based on visual information and potentially textual queries. VideoQA is commonly evaluated using Top-1, Top-K accuracy, and Average Normalized Levenshtein Similarity (ANLS) [91] metric. According to the literature, this task is sub-divided into three sub-categories: *i) Multiple-Choice (MC):* MC-VideoQA addresses multiple-choice question answering. Common benchmarks for this subtask are TGIF-Action and TGIF-Transition [92], MSRVTT-MC [93], and LSMDC-MC [94]. *ii) Open-Ended (OE):* OE-QA answers subjective, creative, and logical questions. Common benchmarks for this subtask are TGIF-Frame [92], MSRVTT-QA [93], MSVD-QA [95], LSMDC-FiB [94], ActivityNet-QA [96]. *iii) Long-Form (LF):* LF-VQA [49] goes beyond single answers, generating comprehensive explanations that understand video content, reason temporally, and adapt to diverse question types. Common benchmarks for this subtasks are ActivityNet-QA [96], How2QA [97] and VIOLIN [98].

Video Captioning. Video captioning generates textual descriptions of video content [33]. MSRVTT [39], Youcook2 [41], and MSVD [99] are the common benchmarks to solve this task. The evaluation metric for this task are BLEU@4 [100], METEOR [101], ROUGE [102], and CIDEr [103].

2.1.3 Video Content Generation and Manipulation

This section covers benchmarks and evaluation metrics for various generative video tasks.

Video Prediction. This area of research encompasses two main sub-tasks: a) *Video future prediction (VFP)*: VFP predicts future frames, given an input video of variable length. Literature uses K600 [24] as a benchmarking dataset and FVD as an evaluation metric for this task; b) *Long-Term Anticipation (LTA)*: LTA [42] predicts next 20 actions given the current action (verb, noun). The common benchmark for this task is Ego4D [104] and the metric to evaluate the performance is *Edit Distance (ED)* [104].

Text-to-video (T2V) Generation. The process [13] involves generating video frames based on a textual prompt. Common

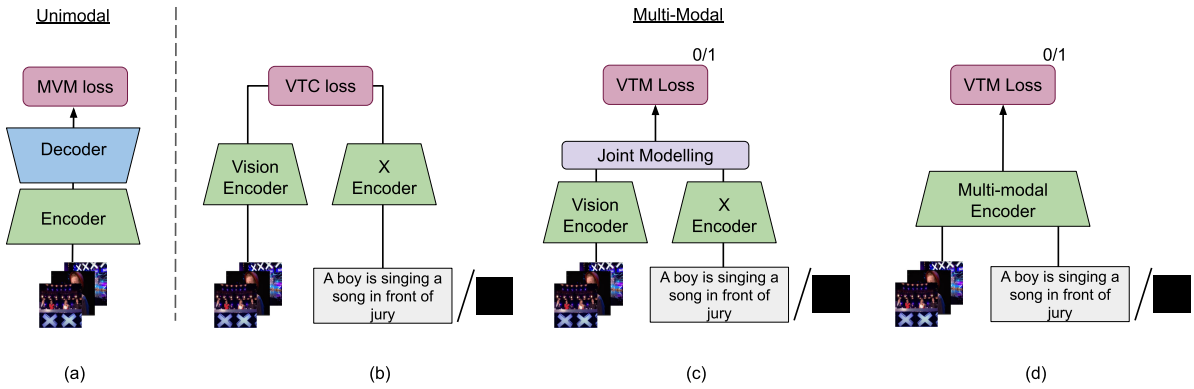


Fig. 4: Figure shows different architectures adopted by Video Foundation Models (ViFMs): Uni-modal ViFMs usually follows (a) Encoder-Decoder (ED) network, and multi-modal foundation model follows either (b) Joint-Encoder (JE) (c) Dual-Encoder (DE) [two input modalities] or Multi-Encoder (MLE) [more than two input modalities], and (d) Mix-Encoder (ME). Best viewed in color.

benchmarks for this task are MSR-VTT [39] and UCF-101 [30], while using Fréchet Video Distance (FVD) [105], CLIP Similarity Score (CLIPSim) [106] and Inception Score (IS) [107] as evaluation metrics.

Video inpainting/outpainting. The task involves predicting the video with the contents filled-in on a masked video using SSv2 [26] as benchmarking dataset and FVD [105] as evaluation metrics.

Video stylization. This task involves generating a video whose style is governed by an additional modality such as text or optical flow [13]. Existing methods try to preserve high-level content of the video, and generating a temporally-consistent stylized version. Benchmark dataset and evaluation metrics for this task are DAVIS 2016 [81] and CLIPSim [106] respectively.

2.2 Architectures and Loss Functions

This section outlines the architectures and training objectives of ViFMs, categorized by modality and loss function, as illustrated in Figure 4.

2.2.1 Architectures

Different architecture patterns in ViFMs can be broadly categorized as: *unimodal* and *multi-modal*. Multi-modal ViFM further show different patterns as discussed below.

Uni-Modal. Unimodal ViFMs [108], [109] focus on a single modality (e.g., video) and typically employ generative objectives like mask reconstruction. They often follow an *Encoder-Decoder* framework where the encoder extracts features from the input video, and the decoder reconstructs the masked or missing parts. The training process is guided by a reconstruction loss function that measures the difference between the original and reconstructed video.

Multi-Modal. Multi-modal ViFMs [110], [111] handle multiple modalities (e.g., video and text) and usually rely on contrastive learning objectives. They typically use encoder-only networks where separate encoders might be employed for each modality. Here, the focus is on learning representations that capture the relationships between different modalities. A contrastive loss function (*VTC*) is used during training to pull together similar representations and push apart dissimilar ones. Within multi-modal models, the encoding style for different modalities can be further classified into three categories (See Fig. 4): i) *Joint-Encoder* (JE) [112] utilizes a single encoder to process all modalities simultaneously. This is computationally efficient but may not capture modality-specific nuances. ii) *Dual/Multi-Encoder* (DE/MLE) [3], [113] employs separate encoders for each modality. While it allows for more specialized feature extraction, it increases computational complexity. Dual encoders are used for two modalities, while architectures handling more than two modalities are referred to as multi-encoders. iii) *Mixed-Encoder* (ME) [114] offers a compromise between joint and separate encoders. It first uses lightweight encoders to extract initial features from individual modalities. These features are then combined and processed by a shared encoder before reaching the final loss function. Both unimodal and multi-modal architectures often leverage transformer blocks as their basic building blocks.

2.2.2 Loss Functions

Video Foundation Models (ViFMs) leverage pretraining with specific objectives, categorized into generative and discriminative tasks. This section explores each category and their associated objectives in detail.

Discriminative. The most common objective functions for ViFMs pretraining are *Video-Text Contrastive (VTC)* [6] and *Video-Text Matching (VTM)* [114]. VTC pulls together similar representations and pushes apart dissimilar ones, while VTM aims to maximize the matching score between a given multimodal pair. These objectives are well-suited for multi-modal architectures but might not be directly applicable to uni-modal architectures. However, contrastive objectives using data augmentation to generate positive pairs have been explored for uni-modal settings [115], [116], [117]. Notably, VIMPAC [118] leverages this approach for ViFMs pretraining.

Beyond VTC and VTM, several variants have been proposed. *Verb-Focused Contrastive (VFC)* [119] focuses on fine-grained verb alignment. *Video-Text Joint (VTJ)* [120] learns a joint representation from video and text (Figure 4 (d) for reference). *Multimodal Temporal Contrastive (MTC)* [49] expands contrastive learning to other modalities, while *Video Clip Contrastive (VCC)* [118] utilizes contrastive learning between video clips. Recent works explore *Tri-Modal Alignment (TMA)*

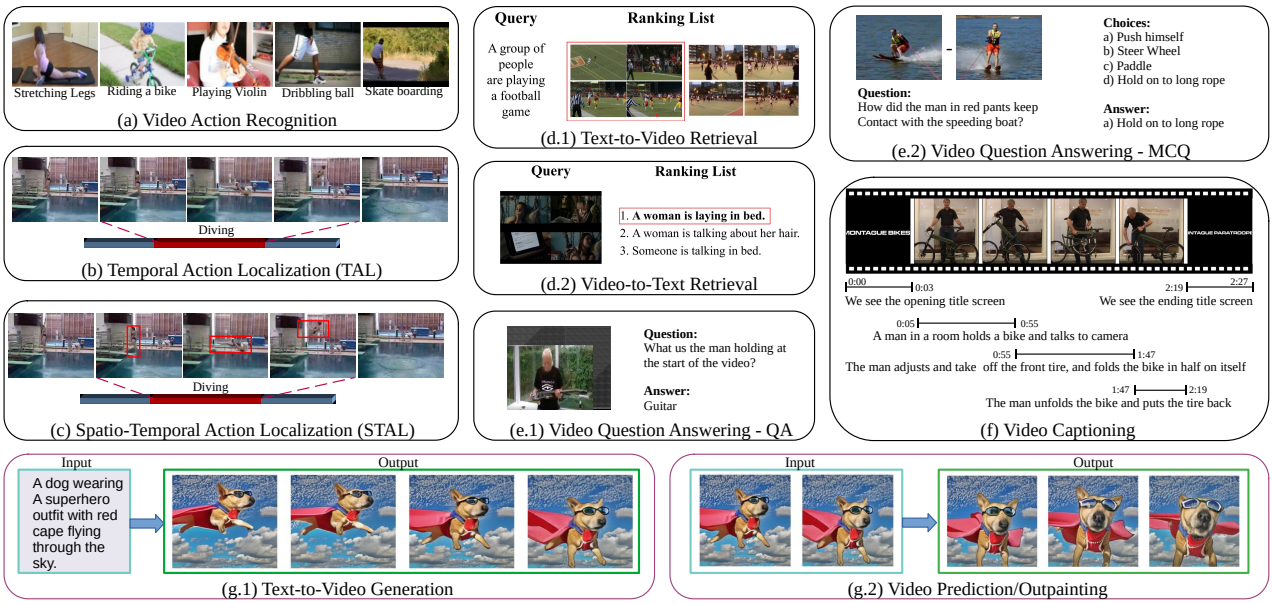


Fig. 5: Figure presents different video tasks: a) Video Action Recognition, b) Temporal Action Localization (TAL), and c) Spatio-temporal Action Localization (STAL), and d) Video-Text Retrieval for video content understanding; e) VideoQA, and f) Video Captioning for descriptive understanding; g) Test-to-video generation, and video prediction for video content generation. Best viewed in color.

[111] for simultaneous cross-modal alignment and fusion, as well as *Omni-Modality Video-Caption Contrastive (OM-VCC)* [113] and *Video-Caption Matching (OM-VCM)* [113] losses. Additionally, *Video-Audio Contrastive (VAC)* [121], [122] and *Attention-Guided Contrastive (AGC)* [123] objectives have been introduced.

The presence of the additional temporal dimension in video allows for significant flexibility in designing discriminative objectives. Several objectives aim to improve temporal modeling capabilities. *Multi-modal temporal relation exploration (MTRE)* [33] and *cross-modal moment exploration (CME)* [33] leverage text guidance to enhance the model's ability to capture temporal context in video. *Time-Order Consistency Check (TOCC)* [124] ensures the correct order of events, while *Control Task (CT)* [124] enforces matching between video events and corresponding text descriptions. *Discriminative Video Dynamics Modeling (DVDM)* [125] specifically promotes nuanced temporal understanding. *Frame-Transcript Matching (FTM)* [126] matches video frames with their corresponding transcripts, and *Temporal Reordering (TR)* [126] predicts the correct order of scrambled video frames.

Generative. This category encompasses various objectives that focus on reconstructing masked information within the video data. Examples include *Mask Language Modeling (MLM)* for text data [127], [128], *Mask Video Modeling (MVM)* for videos [108], [129], *Mask Signal/Data Modeling (MDM, MSM)* for general signals [121], *Mask Frame Modeling (MFM)* for video frames [130], and *Mask Image Modeling (MIM)* for images [131]. The primary goal here is to reconstruct the masked parts, such as predicting a masked frame in MFM. Building upon these core objectives UniVL [120] proposes *Conditional MLM (CMLM)* [120], and *Conditional MFM (CMFM)* [120]. These objectives are primarily used for training unimodal architectures. However, for pretraining multi-modal architectures, generative objectives are often combined with a contrastive loss (acting as a discriminative objective). This combination leverages the strengths of both approaches: reconstructing masked information and learning relationships between modalities.

Beyond the aforementioned objectives, there are less commonly used generative approaches for multi-modal architectures. These include: *Auto-regressive training objectives* like Language Modeling (LM) [33], PrefixLM [33], and next (Image/Motion/Text) [132] token generation, which predicts the next element (token) in a sequence based on the current one. *Captioning Loss* [133] predicts next token based on past video and text. VAST [113] modifies this concept with *Omni-Modality Video Caption Generation (OM-VCG)* loss. *Audio Video Continuation (AVCont)* [13] predicts next frame from audio. Similary, *Video Inpainting and Outpainting*, [13] *Text-to-Video Generation* [13], *Video-to-Text Completion (VTC)* [134], and *Frame Prediction* [13] focus on generating video content.

Task-specific objectives. Beyond general objectives, task-specific objectives are integral to large-scale pretraining. For example, *Prompting Entity Modeling (PEM)* [114] focuses on fine-grained region-entity alignment and action understanding. *Multi-Grained Aligning (MGA)* [135] aligns visual concepts (objects) with text descriptions, while *Multi-Grained Localization (MGL)* [135] locates these concepts in images based on textual descriptions. *Multi-Choice Modeling (MCM)* [136] enhances modality alignment and representation learning. Additionally, *Distillation loss* [137], [138], where a student network mimics the representation of a stronger teacher network, is a common ViFM pretraining objective employed for knowledge transfer, particularly for training lightweight student networks.

2.3 Training Strategy

This section discusses various datasets involved with large-scale pretraining of foundation models. Moreover, this section also briefly mentions the recipe for deploying these models for different video tasks.

2.3.1 Self-supervised Pretraining Datasets

We discuss pretraining datasets for both unimodal and multi-modal architectures in ViFMs in this subsection.

Unimodal. Large-scale models in single modality mostly used a combination of action recognition dataset for self-supervised pretraining. K400 [23], K600 [24], K700 [25], SomethingSomethingV1 (SSv1) [26], and SomethingSomethingV2 (SSv2) [26] are few datasets used for such cases.

Multi-modal. Video-text dataset used for training multi-modal video foundation models are listed as: WebVid-2M [139], HowTo100M [140], EpicKitchen [34], Flinstones [141], Mugen [142]. As we have limited number of multi-modal (Vision-Language) dataset for the video domain. To fulfill this requirement, some foundation models [114], [128], [143], [144], [145], [146] used image-text dataset such CC3M [147], and CC12M [148], and SBU Captions [149] for the pretraining of multi-modal video foundation models. Moreover, few foundation models [49], [113], [127] further curates their own datasets such HD-VILA-100M [127] and LF-VILA-8M [127], and VAST-27M [113] in order to provide diverse and large-scale dataset for multi-modal pretraining.

2.3.2 Semi-supervised Pretraining Datasets

The recent trajectory of multimodal foundation model research reveals a compelling trend toward developing increasingly versatile models. In this subsection, we explore how combining and generating labeled data from multiple sources can enhance the capabilities of ViFMs for a wide range of tasks.

Combining Datasets. The grounding task in the visual domain is not merely solved by self-supervised learning. Preparing large-scale annotated datasets is also a tedious task in such cases. Therefore, different models use a combination of datasets: Object365 [150], OpenImages [151], and COCO [152] for object detection; RefCOCO [153], RefCOCO+ [153], RefCOCOg [83], and VisualGenome [154] for visual grounding; LVIS [155], BDD [69] for tracking; YTVIS19 [79], YTVIS21 [80], RVOS [156] and OVIS [157] for video segmentation.

Pseudo-labelled Datasets. The requirement for large-scale annotated data remains a challenge in computer vision. A recent trend involves leveraging a few powerful teacher models to provide high-quality labels associated with different visual tasks. This approach was pioneered by GRIT [158], which utilizes a teacher model to generate labels for grounding tasks. Following this success, SAM [159] proposes an active learning approach to generate high-quality labeled data specifically for the segmentation task. This approach resulted in the creation of a very large-scale dataset, SA-1B [159], containing 1 Billion high-quality annotations. Similarly, Distill VLM [160] leverages a teacher model to generate captions for existing video datasets like VideoCC [161] and InternVid [162]. This process creates two new pseudo-captioned datasets: VideoCC$^+$, and InternVid$^+$.

2.3.3 Deploying Foundation Models for Video Understanding

Once a ViFM is trained on a large-scale dataset, deploying it for video understanding tasks (Section 2.1) requires further steps. In this subsection, we discuss various approaches to adapt these models for video tasks.

Fine-Tuning. Fine-tuning a model remains a powerful technique for adapting it to specific video tasks [163], [164]. However, its capabilities extend beyond that. Fine-tuning can also be used to improve the model's generic representation, meaning its overall ability to understand and handle various types of information. This type of fine-tuning is often referred to as *post-pretraining* [110], [133], [137], [165] because it essentially trains the model again to expand its knowledge base.

One approach to achieve this effective integration of visual modalities with LLMs is through a process called *instruction-tuning* [166]. This process involves fine-tuning the additional module, or sometimes the entire LLM, on an instructional dataset. The idea behind instruction-tuning was first introduced by InstructBLIP [166]. Instruction has become a common practice for both static images [167], [168], [169] and dynamic videos [170], [171].

Adapters. For video understanding, adapters [172] offer a powerful and efficient approach. These lightweight neural network modules are strategically integrated within large pretrained models. Their key strength lies in requiring training only a limited number of parameters, significantly reducing the computational burden compared to fine-tuning the entire model. This efficiency makes them ideal for video tasks, which often involve processing vast amounts of data. Adapters excel in this domain due to their dual functionality: 1) improving the model's representation for specific tasks [173], [174], [175], [176], [177] and 2) extending its capabilities to enhance the overall understanding of videos [138], [178], [179], [180], [181], effectively creating a more generic video representation.

Prompt-Tuning. Similar to adapter networks, prompt-tuning [182] offers a computationally efficient approach for adapting large pretrained models to new tasks. It achieves this by introducing some additional trainable parameters at the model's input, in the form of a prompt. This prompt essentially guides the pretrained model towards the desired task by providing specific instructions or context. Similar to adapter, they are integrated with large-scale models to improve: 1) representation for specific task [183], [184], [185], and 2) extending the overall video understanding. [186], [187] Alternatively, we can design generic prompts that enhance the model's overall performance on various tasks, e.g., *"image of object"* as used by CLIP [3]. By focusing on a small set of trainable parameters, prompt-tuning significantly reduce the computation complexity.

3 Image-based Video Foundation Models

Image-based Video Foundation Models (ViFMs), the first category in our taxonomy (Fig. 2), adapt image foundation models for video tasks. Unlike the other categories, this subclassification focuses on the inflation approach rather than the training objective. This section explores three primary methods for adapting IFMs for video: post-pretraining (Section 3.1), adapters (Section 3.2), and prompt-tuning (Section 3.3). Each approach can be used to create either a general-purpose ViFM (see Table 1) or a task-specific model.

	Method	Pretraining data		Pretraining Objectives		Architecture		Venue
		Dataset(s)	Size	Discriminative	Generative	Type	Base	
Post-pretraining	VideoCoCa [133]	VideoCC3M [161], HowTo100M [140]	103M	VTC	Captioning	ED	ViT [188], Transformer [189]	arxiv'22
	UMT [110]	K710 [190], WebVid-2M [139], CC3M [147], COCO [152], Visual Genome [154], SBU Captions [149], CC12M [148]	25M	VTM	MLM, MVM	DE	ViT [188], BERT [191], CLIP-ViT [3]	ICCV'23
	MaMMUT [192]	Web alt-text [193]	1.8B	ITC	Captioning	ED	TubeViT [188]	TMLR'23
	Harvest [165]	WebVid-10M [139]	10M	VTC	MLM	ED	UniformerV2 [190], Transformer [189], CLIP [3]	arxiv'23
	CLIP-ViP [194]	WebVid-2.5M [139], HD-VILA-100M [127]	102M	VTC		DE	ViT [188]	ICLR'23
	FitCLIP [137]	WebVid-2.5M [139]	4M	VTC	Distill	DE	CLIP-ViT [3]	BMVC'22
	Distill-VLM [160]	S-Mit [195], WebLI [196]	400K	VTC	Distill	DE	Pali-3 [197], Vit-G [198], UL-2 [199]	CVPR'24
Adapters	EVL [179]	CLIP Pretrained		VTC	-	DE	CLIP [3], Transformer [189]	ECCV'22
	DualPath [200]	CLIP Pretrained		VTC	-	DE	ViT [188], Transformer [189]	CVPR'23
	AG-Adapter [178]	CLIP Pretrained		VTC	-	DE	CLIP [3], LLM [169]	ICCV'23
	DiST [180]	Kinetics-710 [25]	0.5M	VTC		DE	CLIP-ViT [3]	ICCV'23
	RTQ [181]	BLIP Pretrained		VTC, VTM	-	DE	BLIP-ViT [167]	ACMMM'23
	TaCA [138]	LAION-400M [201]	400M	VTC	Distill	DE	ViT [188], BERT [191]	arxiv'23
	PaLM2-VAAapter [202]	WebLI [196], VTP [203], SMiT [195]	-	VTA		DE	CoCa [204], PaLM 2 [205]	arxiv'24
Prompt-Tuning	ATP [206]	CLIP Pretrained		VTC	-	DE	CLIP [3]	CVPR'22
	P-Former [207]	LAION [208], COCO [152], Visual Genome [154], CC-3M [147], SBU Captions [149]	16M	ITC, ITM	ITG	ED	EVA-CLIP [209], LLM [210], Q-Former [211]	NeurIPS'23
	VideoPrompter [187]	CLIP Pretrained		VTC	-	DE	CLIP [3], GPT-3.5 [212], ViFi-Clip [213], AIM [174], Action-CLIP [214]	arxiv'23
Hybrid	Singularity [215]	COCO [152], VG [154], SBU Captions [149], CC3M [147], CC12M [148], WebVid-2M [139]	17M	VTC	MLM	DE	ViT [188], BERT [191]	ACL'23
	PromptCLIP [186]	CLIP Pretrained		VTC	-	DE	CLIP-ViT [3]	ECCV'22
	COSA [216]	CC14M [145], WebVis-2.5M [139]	14M	CITC, CITM	CMLM, CGM	DE	ViT [188], BERT [191]	ICLR'24

TABLE 1: The table categorizes Image-based ViFMs into Post-pretraining, Adapters, Prompt-Tuning, and Hybrid methods. Hyperlinks lead to corresponding code. ◇ Please refer to the supplementary for abbreviations

3.1 Inflation with Post-pretraining

Post-pretraining refines pre-trained IFMs on large video datasets, enhancing their capability for video-centric tasks. We explore both general-purpose ViFMs (detailed in Table 3) in Section 3.1.1 and specialized models in Section 3.1.2 obtained via inflating IFMs.

3.1.1 Generalist Models

In this section, we explore major trends in post-pretraining for video understanding, focusing on leveraging pre-trained IFMs. This approach involves adapting powerful image models to video tasks.

Some approaches directly apply pre-trained models to video frames. For example, *ClipBERT* [217], and *VideoCoCa* [133] employs pre-trained CLIP [3], and contrastive captioner (CoCa) [204] as IFMs respectively. To further optimize IFMs for video tasks, approaches like *Harvest* [165], and *CLIP-ViP* [194] focus on efficient post-training techniques. To elaborate, *CLIP-ViP* [194] for instance proposes auxiliary captions to bridge the domain gap between images and videos while introducing video proxy tokens for efficient processing and Omnisource Cross-modal Learning to jointly learn from diverse data sources.

Student-Teacher Framework for Video Adaptation. A recent trend involves utilizing IFMs as teacher networks, distilling their knowledge to train specialized student networks focused on video understanding. Some approaches adopting this strategy include *FitCLIP*, *Unmasked Teacher (UMT)* [110], and *Distill-VLM* [160]. *Unmasked Teacher (UMT)* [110], and *Distill-VLM* [160] take this approach a step further by implementing stage-wise training. Initially, the student model is trained on video data, followed by distillation loss. Furthermore, *Distill-VLM* also [160] excels in generating high-quality captions for video data. Consequently, it is utilized in producing pseudo-labeled datasets such as VideoCC+ (∼10M Clips) and InternVid+ (∼234M clips). This further improves representation [52], [160] when incorporated during pretraining.

3.1.2 Specialist Models

Post-pretraining can also be tailored to inflate a model specifically for a single video task, akin to fine-tuning IFMs for that task. We discuss example of such approaches in this subsection.

Retrieval and Captioning. Only a few approaches, such as *Clip4Clip* [163] and *Clip4Caption* [164], fine-tune IFMs specifically for retrieval and captioning tasks. Both approaches conduct empirical studies, exploring the efficacy of techniques like feature pooling and fine-tuning on the efficiency of video-text post-training, and its influence on video-text representations for retrieval and captioning tasks.

3.2 Inflation With Adapters

Instead of fine-tuning an entire pre-trained image model, lightweight adapter layers are used for temporal modeling. This adapts the model for video tasks without needing extra pre-training. The section explores methods using adapters to inflate pre-trained image models.

3.2.1 Generalist Models

These adapters integrate with pre-trained image foundation models (IFMs) using only a small number of additional parameters. This allows for efficient generic video representation learning while maintaining the power of pre-trained

models. *DiST* [180] disentangles spatial and temporal learning. This framework leverages a pre-trained image recognition model for spatial understanding and a lightweight temporal encoder to capture dynamic changes, offering an efficient and effective solution for video understanding tasks. *RTQ* [181] clusters redundant tokens and refines redundant information. It further extends the method for complicated video understanding by modeling temporal relations using adapters and querying task-specific information. *EVL* [179] presents a framework that directly trains video models on frozen CLIP [3] features (powerful visual representations learned from image-text pairs). Unlike fine-tuning, EVL uses a lightweight Transformer decoder and a local temporal module to efficiently learn spatio-temporal features without retraining the image backbone. *DualPath* [200] employs two distinct paths: a) Spatial path, which encodes individual frame appearance with minimal tuning, using only a few frames at a low frame rate; and b) Temporal path, which captures dynamic relationships by constructing a grid-like frame set from consecutive low-resolution frames. It then introduces light adapters to both paths for inflating image models for videos. *AGAdapter* [178] proposes two adapter modules: KaAdapter aligns the video-text representation, and PgAdapter employs prompt-tuning to leverage LLMs for captioning. *PaLM2-VAdapter* [202] proposes an adapter module that effectively aligns vision encoders and LLMs using a progressive training strategy, effectively integrating visual information into LLMs. Task-agnostic Compatible Adapter (TaCA) [138] enables seamless integration of new foundation models into existing frameworks without retraining, preserving model strengths through lightweight adapters.

3.2.2 Specialist Models

Adapters also illustrate a wide range of applications for inflating IFMs of specific video tasks. We show some distinguished examples in this section.

Adapters for Text-Video Retrieval. Adapters play a crucial role in text-video retrieval tasks, as demonstrated by various approaches such as PromptSwitch [218], Clip2Video [219], Cross-Modal Adapter [220], AdaCLIP [221], and CrossVTR [222]. *PromptSwitch* [218] introduces a "Prompt Cube" to CLIP's image encoder, capturing global video semantics efficiently, while *CrossVTR* [222] introduces a decoupled video-text cross-attention module to handle spatial and temporal multimodal information separately. *CLIP2Video* [219] simplifies the task into spatial representation and temporal relations, achieving state-of-the-art performance on retrieval benchmarks. Meanwhile, *Cross-Modal Adapter* [220] achieves significant parameter efficiency through adapter-based layers, facilitating realignment of CLIP's feature spaces. *AdaCLIP* [221] addresses the challenge of frame selection and aggregation, offering a tailored system for practical deployment on resource-constrained devices and cloud pipelines.

Adapters for VideoQA. Tem-Adapter [223] tackles VideoQA challenges by bridging domain gaps in image-based pre-trained models. It introduces visual and textual aligners: the "visual temporal aligner" predicts future states based on past video and textual cues, while the "textual semantic aligner" refines textual embeddings using question-answer pairs and video sequences, enabling effective adaptation for VideoQA tasks.

Adapters for Video Action Recognition. While powerful, adapting pre-trained image models for video tasks can be expensive and prone to overfitting. *AIM* [174], *M^2-CLIP* [176], *ST-Adapter* [175], and *ZeroI2V* [177] tackles this challenge using parameter-efficient adapter modules. AIM [174] employs joint spatial-temporal adapters, ST-Adapter [175] uses spatio and temporal adapters, and M^2-CLIP [176] incorporates temporal enhancement and difference modeling to enhance image models with temporal knowledge for action recognition tasks. In contrast, ZeroI2V [177] provides a novel *zero-cost* transfer learning method and proposes a new *spatio-temporal attention* mechanism that captures video dynamics without extra effort.

Adapters for Temporal Action localization (TAL). For zero-shot temporal action detection, *ZEETAD* [224] introduces two key modules: 1) a dual-localization module pinpointing action-relevant regions and generating candidate proposals, and 2) a zero-shot proposal classification module using efficiently fine-tuned CLIP [3] models with adapters for better transferability to the video domain.

3.3 Inflation With Prompt-Tuning

Similar to adapters, prompt tuning improves efficiency by only fine-tuning a few additional parameters. However, these parameters are incorporated at the input layer. This section provides an overview of such approaches.

3.3.1 Generalist Models

Our review highlights the increasing use of prompt-tuning with large language models (LLMs) for video understanding. *Jian* et al. [225] propose a *Prompt-Transformer (P-Former)*, training solely on language data to predict optimal prompts for LLMs. *VideoPrompter* [187] enhances zero-shot performance of existing video-text models by leveraging video-specific information. It employs LLMs with video-specific prompts to generate detailed descriptions and attributes for each class, enriching their textual-representation beyond just the name. A different approach, *Atemporal Probe (ATP)* [206] leverages pre-trained image-language models to extract a single key frame. This key frame then acts as a prompt for a pre-trained vision encoder. Interestingly, their results show that this approach can achieve strong performance on video tasks like question answering and retrieval, suggesting that some video understanding can be achieved by analyzing a single well-chosen frame.

3.3.2 Specialist Models

This subsection explores the approaches inflating IFMs for specific video tasks using prompt-tuning.

Retrieval. In text-video retrieval, *VoP* [183] proposes an efficient adaptation method with minimal parameters by introducing lightweight text and video prompts. These prompts guide the model's learnig towards the relevant inductive-bias while retaining zero-shot capabilities.

Video Action Recognition. *Vita-CLIP* [184] freezes the pre-trained backbone to retain zero-shot capabilities and introduces learnable prompts from both vision and textual

modalities. These prompts capture video-specific information, enhancing representation capabilities.

VideoQA. *Q-ViD* [226] provides instruction prompts to *InstructBLIP* [168] for generating video captions, improving VideoQA. *ViTiS* [185] introduces multimodal prompt learning and a visual mapping network to addresses the challenge of adapting large, pre-trained vision-language models to VideoQA under limited data.

3.4 Inflation With Hybrid Approaches

Some approaches either involve combinations [213], [227] of post-pretraining, adapters, and prompt-tuning or adopting a slightly different method [206], [216] to inflate IFMs into ViFMs. We discuss such approaches in this subsection.

3.4.1 Generalist Models

Some methods explore a combination of different inflating approaches for generic video understanding. *PromptCLIP* [186] efficiently adapts pre-trained models by learning prompt vectors instead of handcrafted ones. These capture video information and act as virtual tokens within the text encoder. PromptCLIP further introduces lightweight transformers (adapter) in the network to capture temporal dynamics. *Singularity* [215] consolidates video clips into single frames using a few additional attention layers, it further modifies the prompts in order to model the temporal relationships that exis in the videos. *COSA* [216] further supports this notion by constructing ”pseudo long-form videos” from existing image-text data. This approach randomly combines multiple images, creating a sequence of static scenes with richer contexts and detailed captions. These ”pseudo videos” allow powerful IFMs, primarily trained on images, to be repurposed for video tasks without explicit temporal modeling.

3.4.2 Specialist Models

We discuss approaches that combine inflating techniques for specific video tasks in this subsection.

Text-To-Video Retrieval. Some approaches utilizes clustering or aggregation techniques to inflate IFMs to video, which is slightly different from the inflation techniques mentioned earlier in this section. CenterCLIP [228], X-Pool [229], and *MEME* [230] presents *clustering-based* approaches for the retrieval task. These approaches bridge the semantic gap between textual queries and video content by grouping similar data points together. *CenterCLIP* [228] identifies the most representative token, *X-Pool* [229] uses text as a condition to guide the aggregation of video tokens, and *MEME* [230] proposes graph patch spreading (GPS) to cluster similar patches together. *ProST* et al. [231] improves the retrieval performance by focusing on fine-grained visual objects (spatial) and interaction (temporal) among them during video-text pretraining and thus inflate IFMs for videos.

Video Action Recognition. Multiple video action recognition approaches build upon IFMs as their foundation. In this paragraph, we explore some examples. For instance, *ViFi-CLIP* [213] first fine-tune CLIP [3] on video data and thus implicitly captures temporal cues without additional modules. It further enhances the performance by learning prompts using a *bridge and prompt* approach in low-data settings. *Wang* et al. [227] captures motion cues through a two-stream adapter block, enriching video representations without sacrificing CLIP's generalization. Additionally, it generates dynamic, motion-aware prompts that describe actions more effectively, guided by captured motion cues. Finally, a pre-matching step aligns video and text representations before feeding them to CLIP, to further boost performance. *BIKE* [232] utilizes a bidirectional knowledge exploration framework (T2V and V2T) from pre-trained IFMs and improves the representation for video recognition.

Adapting SAM [159] for Video Segmentation. Following the success of the SAM [5] for image segmentation, researchers are actively adapting it for video tasks. *SAM-Track* [233] empowers users to interactively segment and track objects through clicks, strokes, or text, while *TAM* [234] achieves high-performance interactive video tracking and segmentation with minimal clicks. For multi-object scenarios, *HQTrack* [235] utilizes SAM for segmentation followed by mask refinement, and *DEVA* [236] incorporates temporal coherency into per-frame SAM segmentation for improved consistency. In the unsupervised realm, *UVOSAM* [237] leverages SAM for video object segmentation without costly annotations. Additionally, *RefSAM* [238] refines SAM for referring video object segmentation (RVOS) by using multi-view information (text, different frames), and *SAM-PT* [239] employs point selection and propagation for zero-shot object tracking and segmentation. These diverse adaptations showcase SAM's potential for various video segmentation tasks, including interactive experiences, leveraging temporal information, and unsupervised learning, thereby marking a significant progress in the field.

3.5 Summary

Image-based ViFMs typically leverage large-scale pretrained image foundation models. These models employ three primary techniques for inflating image models for video tasks: post-pretraining, adapters, and prompt-learning. Some approaches combine one or more of these methods to inflate image models for video tasks. The resulting models can either be used for generic video understanding (Generalist Models), which shows capabilities to generalize across several video tasks; or this inflation is focused towards a specific video task (Specialist Models).

4 Video-based Video Foundation Models

Video-based ViFMs, trained on video/video-X (where X can be text/audio) datasets, and seek to generalize across different video tasks. We classify these models into three primary categories based on their pretraining objectives: generative, discriminative, and hybrid (listed in Table 2). This section explores and discusses major trends in video-based ViFMs.

4.1 Generative Pretraining Objective

Foundation models with a generative objective often employ mask reconstruction as a pretraining objective. The generative objective encompasses MVM, MFM, MIM, and GVM as discussed in Subsec. 2.2.2. In mask modeling, applied masking schemes have evolved with time, transitioning from discrete token masking [118] to random token masking [247],

Method	Pretraining data		Pretraining Objectives		Architecture		Venue
	Dataset(s)	Size	Discriminative	Generative	Type	Base	
Bevt [131]	IN-1K [240], K400 [241]	400M	-	MIM, MVM	ED	Video-Swin [242], VQ-VAE [243]	CVPR'22
ST-MAE [109]	IN-1K [240], K400 [23], K600 [24], K700 [25]	710M	-	MVM	ED	ViT [188]	NeurIPS'22
VideoMAE [12]	K400 [241], SSv2 [26]	400M	-	MVM	ED	ViT [188]	NeurIPS'22
MAM^2 [244]	K400 [241]	400M	-	MVM	ED	ViT [188], VQ-VAE [243]	arxiv'22
MG-MAE [129]	SSv2 [26]	400M	-	MVM	ED	ViT [188]	ICCV'23
AudVis MAE [245]	VGG Sound [246]	0.2M	-	MVM	ED	MAE [247]	ICCV'23
LAVANDER [248]	WebVid-2M [139], CC3M [147]	5M	-	MLM	JE	Video-Swin [242], BERT-base [191]	CVPR'23
MMVG [134]	EpicKitchen [34], Flintstones [141], Mugen [142]	0.5M	-	TVC	JE	VQ-VAE [243], CLIP-Tokenizer [3], VideoSWIN [242]	CVPR'23
MVD [249]	K400 [241]	400M	-	MFM	DD	ViT [188]	CVPR'23
VideoMAEv2 [108]	Unlabeled Hybrid	135M	-	MVM	ED	ViT [188]	CVPR'23
VideoComposer [250]	WebVid-10M [139], LAION-400M [201]	410M	-	CI2VG, CVI, CS2VG	ED	VLDM [251], [252], CLIP-ViT-H [3]	NeurIPS'23
MATS [253]	K400 [241], SSv2 [26], UCF101 [30], HDMB51 [29], Ego4D [104]	>>400M	-	MIM, MVM	ED	ViT [188]	arxiv'23
VideoBERT [254]	Web scraping	300k	-	MLM, MVM	JE	BERT [191], Transformer [189]	ICCV'19
HierVL [42]	Ego4D [104]	3M	VTC	-	Mul-E	Frozen [139], DistillBERT [255]	arxiv'19
TACT [124]	Synthetic Dataset	180M	TOCC, CT	-	-	VideoCLIP	ICLRW'23
VFC [119]	SMiT [256]	0.5M	VTC, VFC	-	DE	PaLM [257], CLIP-ViT [3]	ICCV'23
LAVILA [258]	Ego4D [104], HowTo100M [140]	141M	VTC	-	DE	GPT-2 [259]	CVPR'23
PAXION [125]	ActionBench	0.4M	VTC, DVDM	-	DE	InternVideo, CLIP-ViP, Singularity-temporal	NeurIPS'23
ViCLIP [162]	InterVid [162]	234M	VTC	-	DE	Vit-L [188]	ICLR'24
VideoCLIP [6]	HowTo100M [140]	136M	VTC	-	DE	Vit-L [188], Transfomer [189]	EMNLP'21
UniVL [120]	HowTo100M [140]	136M	VTJ, VTA	CMFM, CMLM, LM	ME	BERT [191], Transformer [189]	arxiv'20
ALPRO [114]	WebVid-2M [139], CC3M [147]	5M	VTC, VTM, PEM	MLM	ME	TimeSformer [11]	CVPR'22
HD-VILA [127]	HD-VILA-100M [127]	103M	VTC	MLM	ME	Bert [191]	CVPR'22
LF-VILA [49]	LF-VILA-8M [49]	8M	VTC, VTM, MTC	MLM	ME	Transformer [189]	NeurIPS'22
TVLT [260]	HowTo100M [140], YTTemporal180M [261]	316M	VAM	MSM	JE	MAE [247]	NeurIPS'22
Vimpac [118]	HowTo100M [140]	136M	VCC	MTP	ED	BERT [191], SimCLR [117]	arxiv'22
SimVTP [262]	WebVid-2M [139]	2M	VTC, VTM	MSM	ED	BERT [191], VideoMAE [12]	arxiv'22
Violet [143]	YT-Temporal [261], WebVid-2.5M [139], CC-3M [147]	186M	VTM	MLM, MVM	ME	Video-Swin [242], LE [263], VQ-VAE [243]	arxiv'22
All-in-One [144]	HowTo100M [140], CC3M [147], WebVid-2.5M [139]	110M	VTM	MLM	JE	ViT [189]	CVPR'23
Hitea [33]	WebVid-2M [139], CC3M [147]	5M	VTC, VTM, MTRE, CME	MLM, PrefixLM	ME	MVit-Base [264], BERT-Base [191]	CVPR'23
Clover [111]	WebVid-2M [139], CC3M [147]	5M	TMA	MLM, MVM	ME	Video-Swin [242], BERT [191]	CVPR'23
VindLu [265]	WebVid-10M [139], CC3M [147], CC12M [148]	25M	VTC, VTM	MLM, MVM	DE	ViT [188], BERT [191]	CVPR'23
VioletV2 [128]	WebVid-2M [139], CC3M [147]	5M	VTM	MLM, MVM	ME	Video-Swin [242], LE [263], VQ-VAE [243], DPT-L [266], RAFT-L [267], SWIN-B [268], DALL-E [269], CLIP-Vit-B [3]	CVPR'23
MuLTI [136]	WebVid-2M [139], CC-3M [147]	5M	VTM, VTC, MCM	MLM	ME	ViT [188], BERT [191]	arxiv'23

TABLE 2: The table showcases Video-based ViFMs, categorized into generative, discriminative, and hybrid pretraining objectives. Hyperlinks direct to corresponding implementations.

with some exploring even intelligent masking strategies [270]. We discuss approaches using different masking schemes in this section.

Discrete Token Masking. Earlier approaches [118], [131], [244] in mask video modeling are based on the prediction of discrete tokens, where each discrete token corresponds to a visual cube from a video. These discrete tokens are generated using dVAE in VQGAN [271] and added into a visual codebook. *Bevt* [131] jointly reconstructs discrete visual tokens within the image and video domains, facilitating the separation of spatial and temporal modeling. *MAM*^{2} [244] propose an encoder-regressor-decoder network followed by two separate decoders to disentangle spatiotemporal modeling. The spatial and temporal decoders in this case reconstruct discrete mask tokens, and RGB difference respectively.

Random Masking. Due to the limitations imposed by the size of the visual codebook, these methods have been replaced by simpler approaches that directly reconstruct masked visual patches. *ST-MAE* [109] extends the concept of Masked Autoencoders (MAE) [247] (for the image domain) to videos, where they propose reconstruction by randomly masking 90% of space-time patches as a challenging pretext task for videos. Different from that, *VideoMAE* [12] considers time as a third independent dimension and proposes masking cubes instead of space-time patches. ST-MAE [109] also observe that randomly masking 90% of video cubes results in effective representation learning. Building upon the VideoMAE [12] framework, *VideoMAEv2* [108] introduces a dual masking strategy that effectively removes cubes from both the encoder and decoder networks, significantly enhancing the model's performance. Additionally, VideoMAEv2 [108] expands its capabilities by incorporating data from multiple sources, further increasing its scale and pre-training efficiency. These approaches incorporate random masking, which might not always result in an optimal representation that can generalize across multiple tasks.

Intelligent Masking Schemes. Some approaches [129], [253] propose intelligent masking schemes, resulting in an effective representation and reducing the computational complexity of the model. *MATS* [253] introduces motion-aware token selection using a pair of adjacent frames. Additionally, this approach introduces motion-aware adaptive frame sampling to further reduce computational complexity. *MGMAE* [183] introduce motion information while masking using optical flow, and thus propose to generate temporally consistent masking/visible volume. Approaches like MVD [249] propose improving the representation by predicting the feature maps instead of raw pixel values. *MVD et al.* [249] propose a dual decoder architecture for efficient spatio-temporal modeling, where one decoder predicts the features of a pretrained image backbone, and the second decoder predicts the features of a pretrained video backbone.

Large Multi-modal Modals (LMMs). Several recent works explore video understanding and generation using LLMs. *ChatVideo* [272] and *MM-VID* [273] convert videos into text for improved comprehension. *VideoChatGPT* [170] and *VideoChat* [274] enhance video-based conversations by integrating visual encoders with LLMs and instruction tuning. *Valley* [275] creates video assistants using curated instruction

datasets and a projection module. *PaLM2-VAdapter* [202] progressively aligns vision and language features using a vision-language adapter module. *VideoDirectorGPT* [276] demonstrates LLMs' potential in video generation tasks with a unique framework. It employs LLMs to plan video content and guide scene-specific video generation, showcasing the versatility of LLMs in both video understanding and creation.

So far, we discuss unimodal approaches and LMMs in this category. *LAVENDER* [110] and *AudVis MAE* [245] are two multimodal approaches based on generative pretraining objective. *LAVENDER* [110] employs text as additional modality and MLM as training objective, whereas *AudVis MAE* [245] proposed unified encoding of audio-visual modalities.

4.2 Discriminative Pretraining Objective

Multi-modal contrastive has emerged as a dominant trend in recent years, surpassing mask reconstruction approaches due to its superior ability to generalize models across different domains. This is because multi-modality incorporates information from multiple sources, such as text and vision, leading to richer and more versatile representations. Research [3], [6] shows that text is the most frequent modality used in conjunction with the visual domain for multi-modal contrastive pretraining. These models are often trained using VTA (discriminative) and VTC (discriminative) as a common pretraining objective. We discuss the research using different variants of discriminative objectives in this section.

Simple Approaches. Approaches like *ViClip* [162] and *VideoClip* [6], aim to create a video counterpart to CLIP [3] (Image-Text Contrastive). These methods rely on video-text contrastive learning as their primary objective. Notably, ViClip [162] further validates the impact of large and diverse training datasets on the quality of learned representations. However, video data poses a greater challenge compared to simpler image-text pretraining due to the additional temporal dimension inherent in video.

Introducing Temporal Consistency and Action Understanding. While simple video-text pretraining struggles with capturing the flow of time in videos, ViFMs like *TACT* [124], *PAXION* [125], *HierVL* [42] and *VFC* [119] offer promising solutions. *TACT* [124] and *PAXION* [125] focus on improving temporal understanding, with TACT enforcing correct event order and PAXION leveraging a knowledge patcher and a specific objective. *HierVL [42]*, on the other hand, aims for comprehensive understanding by analyzing videos at different scales and summarizing both short clips and entire videos. Finally, Verb-focused Contrastive (VFC) [119] excels at capturing fine-grained action details through challenging contrastive examples and precise verb alignment.

Long-form Video Understanding. Long-form video understanding presents challenges due to the memory requirements and model capacities, with only a few attempts extending existing models for this purpose. *LaViLa* [258] investigates how pretrained LLMs can be utilized. This method turns LLMs into "Narrators" by giving them visual inputs, enabling them to automatically create detailed descriptions of long videos. These descriptions are then used to train a video-language model. In a similar direction, *MovieChat* [277] combines ViFMs with LLMs using a Q-former and a projection layer. MovieChat tackles the challenge of processing lengthy videos by introducing a memory management mechanism that reduces complexity and cost while enhancing comprehension.

4.3 Hybrid Pretraining Objective

Hybrid pretraining objectives combine generative objectives, such as mask reconstruction, with discriminative objectives, such as contrastive loss. By integrating both generative and discriminative objectives, these hybrid approaches aim to enhance the learned representations. In this section, we will provide an overview of methodologies that utilize hybrid pre-training objectives.

Simple Approaches. *VIMPAC* [118] is a basic uni-modal approach, which combines a generative task (mask reconstruction) with a contrastive objective (VTC). During contrastive learning, clips from the same video are considered positive pairs, while clips from different videos are considered negative. Conversely, *VideoBERT* [254] represents another straightforward approach, leveraging the robust BERT [191] architecture to accommodate the temporal characteristics of video data.

Advanced Approaches. *UniVL* [120], and *Clover*, [111] uses some advanced approaches. Instead of naively combining objectives, *UniVL* [120] employs stage-by-stage pre-training for both discriminative and generative tasks. On the other hand, *Clover* [111] employs an additional pretraining objective called Tri-Modal Alignment (TMA) (among video, text and video-text joint) to improve cross-modal understanding. Recent advancements extend ViFMs beyond basic video understanding tasks. For example, *MMVG* [134] tackles video storytelling by generating stories from textual prompts. *HD-VILA* [127] focuses on versatility by leveraging a diversely sourced dataset (HD-VILA-100M) for pretraining, enhancing performance across different tasks. Finally, *TVLT* [260] explores understanding multimedia content by focusing solely on raw video and audio, eliminating the need for language.

Improving Action Understanding and Temporal Reasoning. Building on the limitations of basic video-text contrastive pretraining, recent advancements with the hybrid pretraining objective also strive to improve temporal understanding in ViFMs. Approaches like *ALPRO* [114] combine contrastive loss with specialized techniques (e.g., Prompting Entity Modeling (PEM)) for finer-grained video analysis. *Hitea* [33] delves deeper, capturing details of individual moments and their connection to text descriptions through methods like Cross-Modal Moment Exploration (CME). *LF-VILA* [49] tackles long-range dependencies and temporal relationships across modalities with its Multimodal Temporal Contrastive (MTC) and Hierarchical Temporal Window Attention (HTWA) mechanisms. These efforts showcase the ongoing push to strengthen ViFM's ability to grasp the flow of time within videos and extract valuable action knowledge.

Efficient-Effective Approaches. Video Foundation Models (ViFMs) must balance efficiency and performance. ViFMs like *VIOLET* [143] and *VIOLETv2* [128] prioritize complex end-to-end transformer models (e.g., VideoSWIN [242]), leading to higher computational cost. Strategies such as *All-in-One* [144], *SimVTP* [262], and *MuLTI* [136] address this trade-off. *All-in-One* [144] streamlines processing by combining raw video pixels and text tokens in a single model, eliminating

separate encoders. It introduces a "token rolling" operation for effective temporal encoding. *SimVTP* [262] simplifies by using masked autoencoders within a single encoder network, promoting robust video-text representations. *MuLTI* [136] condenses textual features with a "MultiWay-Sampler" for efficient computation and introduces a "Multiple Choice Modeling" pretraining task for enhanced performance. These advances drive the ongoing effort to create more efficient and effective ViFMs for video-language tasks.

Optimizing and Evaluating ViFMs. In contrast to methods focused on specific aspects of pretraining, *VindLU* [265] offers a comprehensive roadmap for effective ViFM pretraining. This work delves into architecture design, fusion techniques, pretraining objectives, data selection, training protocols, and scaling strategies, providing a valuable guide for researchers developing future ViFMs. Furthermore, *MELTR* [278] presents a methodology for fine-tuning VL-FMs to enhance their generalizability across diverse downstream tasks. *VideoGLUE* [279] establishes a standardized evaluation protocol for Video Foundation Models (ViFMs). Finally, *VideoBench* [280] provides a comprehensive benchmark and toolkit, which aims at evaluating the true potential of Video-LMMs towards achieving human-like comprehension and decision-making.

4.4 Summary

Video-based ViFMs train foundational models for video understanding using large-scale video or video-text datasets. We categorize these models based on their types of pretraining objectives: generative, discriminative, and hybrid. Generative pretraining objectives often involve MDM as their primary objective, which is beneficial for spatio-temporal understanding in video. Discriminative objectives typically involve video-text contrastive learning, enhancing semantic understanding as text encapsulates conceptual information. Finally, models with hybrid pretraining objectives achieve the best of both worlds by combining generative and discriminative objectives.

5 Universal Foundational Models

Multi-modal foundation models [52], [294], [296] aim for generalization by integrating additional modalities like audio and sensor data, beyond vision and vision-text modalities, resulting in Universal Foundational Models (UFMs) (last category in our taxonomy). Table 3 lists these approaches. We categorize them based on their pretraining objective (generative, discriminative, hybrid) in this section.

5.1 Generative Pretraining Objective

Similar to video-based ViFMs, UFMs primarily utilize mask reconstruction as a key pretraining objective for generative pretraining. A significant trend in this domain is moving towards unifying architectures and datasets. *VLM* [130] paves the way with a single, streamlined encoder that handles both video and text input in a task-agnostic manner. This simplified architecture, fueled by innovative masking techniques like Mask Modality Modeling (MMM), fosters robust cross-modal understanding without sacrificing individual modality capabilities. *OmniVL* [282] takes unification a step further by proposing a single architecture for both image-language and video-language tasks. This is achieved by first unifying the pretraining datasets and then employing a single encoder for the visual (image and video) domain. Novel contrastive and MLM objectives further support this approach. Consequently, OmniVL excels in both visual-only tasks like image classification and cross-modal tasks like video question answering. *OmniMAE* [283] utilizes masked autoencoding to train a single encoder-decoder network for both images and videos. This approach could be easily generalized to other visual modalities such as thermal images and 3D point clouds. Noteworthy is OmniVORE [112], which combines various visual modalities in a similar manner but for supervised classification tasks. A different approach, *MaskFeat* [281] focuses on redefining the pretext task as Mask Feature Prediction (MFP) to design a unified architecture for both image and video understanding, where features in this case are HOG (Histogram of Oriented Gradients).

LLM-based Approaches. Recent advancements have integrated multiple modalities, including image, video, text, and audio, into LLMs, leading to the development of general-purpose LMMs. Some approaches focus on combining different visual modalities such as image and video. *VideoLaVIT* [132] breaks down videos into keyframes and temporal motions, enabling unified pretraining across diverse modalities, including images and videos. Moving towards image-video understanding, *VideoLLaVA* [304] aligns visual representations from images and videos before projecting them onto the LLM, refining them through instruction tuning. Additionally, *Chat-UniVi* [305] proposes a unified approach for image and video understanding using dynamic visual tokens and a multi-scale architecture to efficiently represent and perceive semantics and details simultaneously.

On the other hand, some approaches focus on integrating the audio modality along with the visual modalities into LLMs. *Video-LLaMA* [7] pioneers the integration of visual and auditory information through separate Q-formers and pretrained encoders. *FAVOR* [306] tailors a framework for audio-visual LLMs, incorporating a "Causal Q-Former" that considers causal relationships between video frames.

Meanwhile, the trend is now slightly moving towards combining all modalities, including image, video, audio, and text, using a single foundational model. *Macaw-LLM* [307] directly integrates visual, audio, and textual features, facilitating a unified understanding of videos. Conversely, *VideoPoet* [13] utilizes a decoder-only transformer architecture similar to LLMs for generating high-quality videos with matching audio based on textual input, particularly excelling in "zero-shot" scenarios. Advancements continue towards audio-visual grounding, with PG-Video-LLaVa [308] enhancing LLMs for video comprehension and object grounding by introducing pixel grounding capabilities through object tracking and audio transcription.

Large-scale Models for Generative Tasks. With the advent of large-scale pretraining, two main lines of video generation approaches have emerged: autoregressive transformers [13], [250], [276], [309], [310], [311], [312] and diffusion models [313], [314], [315], [316], [317], [318], [319]. Autoregressive transformers [320] generate sequences (like text or video) one element at a time, considering previously generated elements

Method	Pretraining data		Pretraining Objectives		Architecture		Venue
	Dataset(s)	Size	Discriminative	Generative	Type	Base	
VLM [130]	HowTo100M [140]	1.1M	-	MFM, MLM, MMM	JE	BERT [191]	ACL'21
MaskFeat [281]	IN-21K [240], K400 [23]	400M	-	MFP	ED	MViT [264]	CVPR'22
OmniVL [282]	IN-1K [240], Something-Something V2 [26],	1.4M	-	MDM	JE	BERT [191], TimeSformer [11]	NeurIPS'22
OmniMAE [283]	IN-1K [240], Something-Something V2 [26],	1.4M	-	MDM	JE	ViT [188]	CVPR'23
Video-LaVIT [132]	WebVid-10M [139], CC3M [147], CC12M [148], SBU-Captions [149], BLIP-Capfilt [167], RedPajama [284], Instructional data [170], [285]	103M	-	N(I/M/T)TG	JE	EVA-CLIP-ViTg [209]	arxiv'24
VideoPoet [13]	Web Scraping	1.25B	-	T2V, T2I, FP, Central inpainting and outpainting, AVCont	MLE	LLM [286]	arxiv'24
VATT [122]	HowTo100M [140], AudioSet [287]	27M	VAC, VTC	-	MLE	Tranfomers [189]	NeurIPS'21
VITO [123]	K400 [241]	400M	AGC	-	ED	ViT [188], VQ-VAE [243]	arxiv'22
LanguageBind [288]	VIDAL-10M	10M	MMC	-	MLE	Open-CLIP [289]	ICLR'24
X^2VLM [135]	COCO [152], Visual-Gnome [154], SBU Captions [149], CC [147], Object365 [150], OpenImages [151], WebVid-2.5M [139], HowTo100M [140], Yt-Temporal [261]	28M	MGA, MGL	-	ME	Transformer [189]	TPAMI'24
MERLOT [126]	YT-Temporal-180M [261]	6M	FTM, TR	MLM	DE	ViT [189], RoBERTa [290]	NeurIPS'21
InternVideo [291]	Kinetics-400 [23], WebVid-2M [139], WebVid-10M [139], HowTo100M [140], AVA [60], Something-Something V2 [26], Kinetics-710 [25]	12M	VTC	MVM	DE	ViT [189], UniformerV2 [190]	arxiv'22
ViC-MAE [292]	MiT [293], K400 [241]	1.2M	ITC	MIM	DE	ViT [188]	arxiv'22
Perciever-VL [146]	CC3M [147], WebVid-2.5M [139]	5M	VTM	MLM	JE	ViT [188], BERT [191]	WACV'23
Smaug [294]	COCO [152], Visual-Gnome [154], SBU Captions [149], CC3M [147], CC12M [148], WebVid-2M [139]	17M	VTM, VTC	MVM, MLM	ME	ViT [188]	ICCV'23
CAV-MAE [121]	AudioSet2M [287]	2M	VAC	MSM	ME	ViT [189]	ICLR'23
VAST [113]	VAST-27M [113]	27M	OM-VCC, OM-VCM	OM-VCG	ME	BERT [191], BEAT [295], EVAClip-ViT-G [209]	NeurIPS'23
mPLUG-2 [296]	MS-COCO [152], Visual Genome [154], CC12M [147], SBU Captions [297], WebVid-2M [139], WikiCorpus [191], Crawled data	30M	VTC, VTM	MLM	ME	BERT [191], Transformer [189]	ICML'23
VALOR [145]	VALOR-1M [147], WebVid-2.5M [139], CC14M [147], HD_VILA_100M [127]	119M	MGA	MGC	MLE	BERT [191], CLIP [3], VideoSwin [242], AST [298]	arxiv'23
InternVideo2 [299]	K-Mash [299], MVid [299], WebVid [139], InternVid [52]	6B	Con, VLAA	MLM, Distill	MLE	ViT [188], BERT [191], BEATs [295]	arxiv'24
VideoPrism [300]	Anonymous Corpus [300]	1.3B	VTC	MVM, distill	JE	ViT [188], ViViT [17]	arxiv'24
GLEE [301]	Object365 [150], OpenImages [151], COCO [152], LVIS [155], BDD [69], YTVIS19 [79], YTVIS21 [80], OVIS [157], RefCOCO [153], RefCOCO+ [153], RefCOCOg [83], VisualGenome [154], RVOS [84], SA1B [159], GRIT [158]	5M	SL, BL, ML, CL, CTL	Distill	MLE	ResNet-50 [302], Swin-L [268], EVA-02-L [303]	CVPR'24

TABLE 3: Table shows UFMs integrating multiple modalities beyond video and text. It categorizes methods into generative, discriminative, and hybrid pretraining objectives, with hyperlinks to corresponding implementations.

to predict the next. Some autoregressive approaches, like SORA [311], VideoPoet [13], and VideoDirectorGPT [276], demonstrate the ability to generalize across multiple video tasks (e.g., VideoQA, Video Generation) by incorporating autoregressive language models into their architectures. Diffusion models [251], [252], on the other hand, gradually add noise to a video sample and then learn to reverse the process to synthesize videos from this noise. Large-scale diffusion models achieve impressive results on specific video generation tasks such as video editing [316], [321], [322], video synthesis [323], [324], and text-to-video generation [313], [314], [315]. These approaches could potentially inspire improvements in temporal reasoning within ViFMs.

5.2 Discriminative Pretraining Objective

Universal models aim to achieve comprehensive understanding by processing different modalities (e.g., image, video, audio, text) together. *VITO* specifically targets the fusion of image and video modalities through attention-guided contrastive learning (AGC) and harnesses a large-scale video dataset known as *VideoNet*, akin to ImageNet but tailored for videos. Similarly, *VATT* [122] adopts a strategy of projecting and aligning different modalities—audio, video, and text—using a cross-modal encoder to facilitate multi-modal comprehension. In contrast, X^2VLM [135] proposes a modular architecture that offers the flexibility to integrate additional modalities seamlessly without necessitating the retraining of the entire framework. Meanwhile, *LanguageBind* [288] addresses the expansion of modalities by leveraging language as a central anchor. It employs a pretrained video-language model, preserving its language encoder while training new encoders for supplementary modalities like audio or depth through contrastive learning with a *multi-modal contrastive (MMC)* objective. This process aligns all modalities within a shared feature space, enhancing the model's overall understanding. Notably, for object-centric video tasks, existing ViFMs may not be suitable. Recently, the introduction of GLEE [301], an object-centric foundation model, extends the scope of research in ViFMs for video tasks by incorporating visual prompts alongside vision-text input.

5.3 Hybrid Pretraining Objective

While generative pretraining objectives like mask modeling enhance the spatio-temporal understanding of videos, multi-modal contrastive learning improves the semantic understanding. Hybrid approaches aim to achieve the best of both worlds by combining these techniques during pretraining. *MERLOT* [126] exemplifies this by employing Frame-Transcript Matching (FTM) and Temporal Reordering (TR) to align video frames with their captions, alongside MLM objective for deeper language grasp. *InternVideo* [291] takes a similar approach, leveraging MVM objective to capture video actions and VTC objective to create a shared semantic space for video and language. It further strengthens this representation with supervised action classification and cross-modal attention. Finally, *ViC-MAE* [292] utilizes MAE [247] to capture local features in video patches for fine-grained understanding. It then employs contrastive learning

and pooling across video frames to extract global features representing the entire video.

Efficient Approaches. Efficiency in pretraining video foundation models (ViFMs) is critical due to their computational expensiveness, especially with multiple or hybrid objectives. To tackle this challenge, researcher incorporates specialized attention mechanisms and strategies. For instance, *Perceiver-VL* [146] utilizes iterative latent attention, a technique that bypasses the computational bottleneck of standard self-attention in transformers, leading to significant efficiency gains. Additionally, *Smaug* [294] leverages MAE for efficient pretraining, masking both visual and textual inputs to reduce costs and improve cross-modal alignment. It further employs a space-time token sparsification module to strategically select informative regions and frames, minimizing computational demands.

Effective Approaches. Recent advancements have pushed the boundaries of ViFMs by incorporating multiple modalities beyond just video. *VAST* [113] pioneers an "omni-modal symphony" by incorporating vision, audio, subtitles, and text, leveraging the VAST-27M dataset for comprehensive multimodal training. Similarly, *CAV-MAE* [121] extends the MAE paradigm to video, introducing audio reconstruction during pretraining and integrating both masked modeling and contrastive objectives for enhanced comprehension. Building on prior work, *InterVideo2* [299] introduces the audio modality and a progressive training approach to generalize across multiple video and audio interaction tasks. Pushing the boundaries of multimodal learning, *VALOR* [145] proposes vision-audio-language omni-perception models with discriminative and generative pretraining tasks, facilitating cross-modal connections and empowering the model for diverse tasks like retrieval and captioning. Meanwhile, *VideoPrism* [300] adopts a two-stage training process, refining spatio-temporal representations with video-text data and employing techniques like global-local distillation, yielding versatile representations for varied video understanding tasks.

Modular architectures, exemplified by X^2VLM [135] and *mPLUG-2* [296], further enhance ViFMs flexibility, with separate encoders for each modality and shared attention and contrastive learning modules, enabling tailored models for specific tasks and improved transferability across domains. This modular approach fosters collaboration while addressing the evolving needs of multimodal understanding in ViFMs.

5.4 Summary

Universal Foundation Models (UFMs) aim to integrate multiple modalities (apart from video and text), such as images and audio, into a single cohesive framework for video understanding tasks. Typically, this additional modality is either images—treating images and videos as distinct modalities to unify different visual types—or audio, which directly complements video content. When comparing results from different categories of foundational models, we observe that UFMs outperform others on various video tasks by incorporating these additional modalities. Similar to video-based ViFMs, we categorize UFMs based on their types of pretraining objective types.

	Method	Arch. Type	K400 [23]		HMDB51 [29]		UCF101 [30]		SSv2 [26]	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Generalist	VideoCoca [133]	ED	72.0	90.5	58.7	84.5	**86.6**	**98.4**	-	-
	EVL [179]	DE	82.9	-	-	-	-	-	61.7	-
	DualPath [200]	DE	85.4	97.1	-	-	-	-	70.3	92.9
	UMT-B [110]	DE	87.4	97.5	-	-	-	-	70.8	92.6
	UMT-L [110]	DE	90.6	**98.7**	-	-	-	-	74.7	94.7
	All-in-One [144]	JE	53.2	83.5	55.2	**89.1**	84.1	95.7	-	-
	BEVT [131]	ED	80.6	-	-	-	-	-	70.6	-
	MAM^2 [244]	ED	85.3	96.7	-	-	-	-	71.3	93.1
	MG-MAE [129]	ED	81.8	95.0	-	-	-	-	72.3	93.5
	ST-MAE [109]	ED	81.3	94.9	-	-	-	-	72.1	93.9
	VATT [122]	JE	79.9	94.6	-	-	-	-	-	-
	VideoMAE-B [12]	ED	81.5	95.1	-	-	-	-	70.8	92.4
	VideoMAE-H [12]	ED	86.1	97.3	-	-	-	-	75.4	95.2
	VideoMAEv2-H [108]	ED	88.6	97.9	-	-	-	-	76.8	95.8
	VideoMAEv2-g [108]	ED	90.0	98.4	-	-	-	-	77.0	**95.9**
	VIMPAC [118]	Enc	77.4	-	-	-	-	-	68.1	-
	InternVideo [291]	DE	**91.1**	-	**89.3**	-	-	-	**77.2**	-
	mPLUG-2 [296]	ME	87.1	97.7	-	-	-	-	-	-
	MaskFeat-S [281]	ED	82.2	95.1	-	-	-	-	-	-
	MaskFeat-L [281]	ED	86.4	97.1	-	-	-	-	74.4	94.6
	OmniMAE [283]	JE	84.0	-	-	-	-	-	69.5	-
	OmniVL [282]	JE	79.1	94.5	-	-	-	-	-	-
	VideoPrism-B [300]	JE	84.2	-	-	-	-	63.6	-	-
	VideoPrism-g [300]	JE	87.2	-	-	-	-	-	68.5	-
Specialist	AIM [174]	DE	84.7	96.7	-	-	-	-	69.1	92.2
	ActionCLIP [214]	DE	83.8	97.1	-	-	-	-	-	-
	BIKE [232]	DE	88.6	98.3	-	-	-	-	-	-
	Frozen [139]	DE	50.5	80.7	54.3	88.0	81.3	94.3	-	-
	MIL-NCE [325]	DE	-	-	53.1	87.2	82.7	-	-	-
	M^2-CLIP [176]	DE	84.1	96.8	-	-	-	-	69.1	91.8
	ST-Adapter [175]	DE	82.7	96.2	-	-	-	-	69.5	92.6
	Vita-CLIP [326]	DE	82.9	96.3	-	-	-	-	48.7	-
	X-CLIP [327]	DE	87.7	97.4	-	-	-	-	-	-

TABLE 4: Comparing the finetuned performance of state-of-the-art (SOTA) generalist (Image-based, Video-based, UFMs) and specialist models for **video action recognition** on K400 [23], HMDB51 [29], UCF101 [30], and SSv2 [26] datasets. The highlighted entries indicate the best performing methods.

6 Comparison and Discussion

We compare state-of-the-art (SOTA) performance on six video understanding tasks: action recognition (zero-shot and open-vocabulary), text-to-video retrieval, action localization, video question answering (VideoQA), video captioning, and text-to-video generation (found in Tables 4, 5, 6, 7, 8, and 9). Each table categorizes models as either generalist — capable of handling multiple tasks (further subdivided into image-based, video-based, and universal foundational models (UFMs), as discussed in previous sections), or specialist — excelling at a single task.

6.1 Video Content Understanding

Video action recognition, text-to-video retrieval, and spatio-temporal action localization are established tasks for video content understanding in the literature. This section compares the performance of state-of-the-art foundation models across these tasks.

6.1.1 Video Action Recognition

Table 4 compares Top-1 and Top-5 Accuracy of foundation models on video action recognition. Universal foundational model (UFM), i.e., InternVideo [291] consistently achieves top-1 accuracy on K400 [23], HMDB51 [29], and SSv2 [26]. For top-5 accuracy, UMT-L [110] excels on K400, All-in-one [144] on HMDB51, and VideoMAEv2-g [108] on SSv2. Notably, VideoCoCa [133], an image-based ViFM, leads in both top-1 and top-5 accuracy on UCF101 [30]. This variety

Method	Arch. Type	HMDB51 Top-1	UCF101 Top-1	K400 HM	HMDB51 HM	UCF101 HM	SSv2 HM
		Zero-Shot		Base-To-Novel			
ActionCLIP [214]	DE	40.8	58.3	52.6	48.5	70.7	11.5
AIM [174]	DE	-	-	68.0	57.1	82.6	8.2
BIKE [232]	DE	52.8	80.8	-	-	-	-
EVA-CLIP [209]	DE	-	76.8	-	-	-	-
EZ-CLIP [328]	DE	55.2	82.6	66.3	**66.3**	85.4	**14.8**
FitCLIP [137]	DE	-	73.3	-	-	-	-
Froster [329]	DE	-	-	**70.4**	65.1	**87.0**	14.6
IMP [330]	ME	**59.1**	**91.5**	-	-	-	-
LSS [331]	DE	51.4	74.2	-	-	-	-
M^2-CLIP [176]	DE	47.1	78.7	-	-	-	-
MAXI [332]	DE	52.3	78.2	-	-	-	-
MOV [333]	DE	57.8	80.9	-	-	-	-
PromptCLIP [334]	DE	-	66.6	-	-	-	-
St-Adapter [175]	DE	-	-	67.3	55.9	80.9	8.8
ViFi-CLIP [213]	DE	51.3	76.8	67.9	61.9	78.3	13.9
Vita-CLIP [326]	DE	48.6	75.0	-	-	-	-
X-CLIP [327]	DE	44.6	72.0	64.0	55.0	71.2	7.4

TABLE 5: Comparing the zero-shot and base-to-novel generalization (or open-vocabulary) action recognition performance of SOTA ViFMs for video action recognition on K400 [23], HMDB51 [29], UCF101 [30], and SSv2 [26] datasets. The highlighted entries indicate the best performing methods in both zero-shot and base-to-novel generalization settings.

underscores the strengths of unified, image-based, and video-based models in video action recognition, each excelling based on the benchmark and evaluation metric. However, universal foundation models tend to perform the best in most cases.

Zero-shot and Open-vocabulary Action Recognition. Recent architectures that include multiple modalities extend video action recognition to zero-shot and open-vocabulary tasks, using language to help understand unseen classes. Table 5 (left) compares top-1 accuracy of ViFMs for zero-shot experiments on HMDB51 [29] and UCF101 [30]. In zero-shot results, IMP [330] (a UFM) performs best on both datasets. Table 5 (right) shows the harmonic mean (HM) of base and novel class accuracies for open-vocabulary tasks on Kinetic400 [23], HMDB51 [29], UCF101 [30] and SSv2 [26]. Froster [329] leads on K400 [23] and UCF101 [30], while EZ-Clip [328] (image-based) excels on JHMDB [29] and SSv2 [26]. Lower performance in most cases, especially on the complex SSv2 [26] dataset, indicates significant room for improvement in multi-modal action recognition.

Discussion. ViFMs perform well on large datasets like Kinetics400 [23] and SSv2 [26], but face challenges when transferring to smaller datasets like HMDB51 [29] and UCF101 [30]. The size disparity complicates fine-tuning, resulting in few reported results. Additionally, zero-shot and base-to-novel performance across benchmarks needs improvement. Enhancing representational spaces to better capture spatio-temporal context and vision-language semantics is crucial for progress.

Few foundation models, like LaViLa [49] and Avion [341], target action recognition on complex datasets like EpicKitchen [34] and Ego-4D [104], which feature longer, egocentric videos. This differs significantly from the third-person videos used in pretraining, posing a unique challenge. Even generic models like VideoPrism [300] struggle with this view translation, underscoring the need for models specifically designed for such complexities.

	Method	Arch. Type	MSR-VTT [39] R@1	R@5	R@10	DiDeMo [40] R@1	R@5	R@10	LSMDC [19] R@1	R@5	R@10
	Fine-Tune										
Generalist	PromptCLIP [334]	DE	36.5	64.6	-	36.1	64.8	-	13.4	29.5	-
	RTQ [181]	DE	53.4	76.1	84.4	57.6	84.1	89.8	-	-	-
	Singularity [215]	DE	36.8	65.9	75.5	47.4	75.2	84.0	-	-	-
	VideoCoca [133]	ED	34.3	57.8	67.0	-	-	-	-	-	-
	UMT-B [110]	DE	51.0	76.5	84.2	61.6	86.8	91.5	32.7	54.7	63.4
	UMT-L [110]	DE	58.8	81.0	87.1	70.4	90.1	93.5	43.0	65.5	73.0
	All-in-One [144]	JE	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-
	ALPRO [114]	ME	33.9	60.7	73.2	35.9	67.5	78.8	-	-	
	Clover [111]	ME	40.5	69.8	79.4	50.1	76.7	85.6	24.8	44.0	54.5
	HD-VILA [127]	ME	35.6	65.3	78.0	28.8	57.4	69.1	17.4	34.1	44.1
	Hitea [33]	ME	44.4	69.3	78.9	51.8	79.1	85.3	27.1	46.2	54.5
	LAVANDER [248]	JE	37.8	63.8	75.0	47.4	74.7	82.4	22.2	43.8	53.5
	SimVTP [262]	ED	53.6	81.9	**90.7**	-	-	-	-	-	-
	Smaug [294]	ME	44.0	70.4	78.8	55.6	80.8	88.4	-	-	-
	VideoCLIP [6]	DE	30.9	55.4	66.8	-	-	-	-	-	-
	VindLU-L [265]	DE	48.8	72.4	82.2	59.8	86.6	**91.5**	-	-	-
	VIOLET [143]	ME	34.5	63.0	73.4	32.6	62.8	74.7	16.1	36.6	41.2
	InternVideo [291]	DE	55.2	79.6	87.5	57.9	82.4	88.9	34.0	53.7	62.9
	mPLUG-2 [296]	ME	53.1	77.6	84.7	56.4	79.1	85.2	**34.4**	**55.2**	**65.1**
	OmniVL [282]	DE	47.8	74.2	83.8	52.4	79.5	85.4	-	-	-
	X^{2}VLM [135]	ME	49.6	76.7	84.2	-	-	-	-	-	-
	VALOR [145]	MLE	54.4	79.8	87.6	57.6	83.3	88.8	31.8	52.8	62.4
	VAST [113]	ME	**63.9**	**84.3**	89.6	**72.0**	**89.0**	91.4	-	-	-
Specialist	CAMoE [335]	DE	47.3	74.2	84.5	43.8	71.4	79.9	25.9	46.1	53.7
	Clip4Clip [163]	DE	42.1	71.9	81.4	43.4	70.2	80.6	21.6	41.8	49.8
	CrossTVR [222]	ME	54.0	77.5	85.3	55.0	77.6	-	27.7	48.5	-
	Frozen [139]	DE	32.5	61.5	71.2	31.0	59.8	72.4	15.0	30.8	39.8
	MCQ [336]	DE	37.6	64.8	75.1	37.0	62.2	73.9	17.9	35.4	44.5
	MILES [337]	DE	37.7	63.6	73.8	36.6	63.9	74.0	17.8	35.6	44.1
	ProST [231]	DE	46.9	73.3	82.9	47.5	75.5	84.4	26.3	46.1	55.2
	OA-Trans [338]	DE	35.8	63.4	76.5	34.8	64.4	75.1	18.2	34.3	43.7
	QB-Norm [339]	-	47.2	73.0	83.0	43.3	71.4	80.8	22.4	40.1	49.5
	TMVM [340]	DE	36.2	64.2	75.7	36.5	64.9	75.4	17.8	37.1	45.9
	X-CLIP [327]	DE	46.1	73.0	83.1	45.2	74.0	-	23.3	43.0	-
	Zero-Shot										
Generalist & Specialist	MCQ [336]	DE	26.0	46.4	56.4	25.6	50.6	61.1	12.2	25.9	32.2
	MILES [337]	DE	26.1	47.2	56.9	27.2	50.3	63.6	11.1	24.7	30.6
	OA-Trans [338]	DE	23.4	47.5	55.6	23.5	50.4	59.8	-	-	-
	UMT-B	DE	35.2	57.8	66.0	41.2	65.4	74.9	19.1	33.4	42.2
	UMT-L	DE	40.7	63.4	71.8	48.6	72.9	79.0	**24.9**	**41.7**	**51.8**
	FitCLIP [137]	DE	33.8	59.8	69.4	28.5	53.7	64.0	-	-	-
	Frozen [139]	DE	18.7	39.6	51.6	21.1	46.0	56.2	9.3	22.0	30.1
	ALPRO [114]	ME	24.1	44.7	55.4	23.8	47.3	57.9	-	-	-
	Clover [111]	ME	26.4	49.5	60.0	29.5	55.2	66.3	14.7	29.2	38.2
	VideoCLIP [6]	DE	10.4	22.2	30.0	16.6	46.9	-	-	-	-
	UniVL [120]	ME	21.2	49.6	63.1	-	-	-	-	-	-
	VIOLET [143]	ME	25.9	49.5	59.7	23.5	49.8	59.8	-	-	-
	InternVideo [291]	ME	40.0	65.3	**74.1**	31.5	57.6	68.2	17.6	32.4	40.2
	OmniVL [282]	DE	34.6	58.4	66.6	33.3	58.7	68.5	-	-	-
	VAST [113]	ME	**49.3**	**68.3**	73.9	**55.5**	**74.3**	**79.6**	-	-	-

TABLE 6: Comparing the fine-tuned and zero-shot performance of SOTA generalist (Image-based, Video-based, UFMs) and specialist models for **text-to-video retrieval** tasks on MSR-VTT [39], DiDeMo [40], and LSMDC [19] dataset. The highlighted indicate the best performing methods in both fine-tuned and zero-shot settings.

6.1.2 Text-to-Video (T2V) Retrieval.

Table 6 compares rank-1/5/10 accuracy of SOTA foundational models for text-to-video (T2V) retrieval. Unified models incorporating video, text, and potentially audio modalities generally perform better across various datasets. For example, VAST [113] excels on MSR-VTT [39] and DiDeMo [40], while mPlug-2 [296], focused on unified image-video, dominates on LSMDC [19] under fine-tuned settings. In zero-shot scenarios, VAST [113] leads on MSR-VTT [39] and DiDeMo [40], whereas UMT [110], built by inflating image models, performs best on LSMDC [19]. This indicates that datasets like LSMDC [19], with movie descriptions, benefit from the strong spatio-temporal understanding provided by UMT (MDM objective) and mPlug-2 (temporal module).

Discussion. While ViFMs have made significant progress in recognition tasks, text-to-video (T2V) retrieval remains

Method	TAL		STAL	
	ANet	THUMOS14	AVA	AVA-Kinetics
MaskFeat-L [281]	-	-	37.8	-
ST-MAE-L [109]	-	-	37.3	-
VideoMAE-L [12]	-	-	39.3	-
VideoMAEv2 [108]	-	69.6	**42.6**	**43.9**
UMT-B [110]	-	-	33.5	-
UMT-L [110]	-	-	39.8	-
InternVideo [162]	**39.0**	**71.6**	41.0	42.5
VideoPrism-B [300]	36.6	-	30.6	31.8
VideoPrism-g [300]	37.8	-	36.2	37.3

TABLE 7: Comparing the fine-tuned performance of SOTA Vi-FMs on ANet/ActivityNet [27] for Temporal Action Localization (TAL) and AVA [60] and AVA-Kinetics [61] for Spatio-temporal action localization (STAL). The highlighted entries indicate the best performing methods.

challenging, as evidenced by lower top-1 and top-5 accuracy across benchmarks. This task requires effective multi-modal architectures to capture the interaction between video content and textual descriptions. Aligning video content with text is difficult due to the complex structures in videos, including spatio-temporal aspects, compared to the simpler structures of language. Results indicate that incorporating additional modalities, such as audio, is beneficial for T2V tasks.

Moreover, T2V retrieval is problematic for long-term video understanding, as seen in the LSMDC [19] benchmark with its long movie videos. Additionally, new challenges like Boundary Caption-Video Retrieval, as described in Kinetic-GEBC [342], which aims to retrieve videos containing specific boundaries based on their descriptions, require further exploration.

6.1.3 Spatio-temporal Video Understanding

Table 7 compares the mean Average Precision (mAP) of various foundation models on Temporal Action Localization (TAL) and Spatio-temporal Action Localization (STAL) tasks. TAL evaluation uses the ActivityNet [27] and THUMOS14 [53] datasets, while STAL evaluation employs the AVA [60] and AVA-Kinetics [61] datasets. InternVideo [52] excels in TAL, and VideoMAEv2 [108] leads in STAL.

Discussion. Table 7 reveals that most models use MDM for pretraining, effectively handling action localization but struggling with semantic understanding. Further research is needed in spatio-temporal modeling for ViFMs. Combining text and integrating generative and discriminative objectives are crucial for comprehensive understanding.

While ViFMs demonstrate strong performance, they haven't been evaluated on additional datasets like UCF24 [30] and JHMDB [29], likely due to their smaller scale limiting fine-tuning effectiveness. High-resolution datasets like UCF-MAMA [59] and VIRAT [31], with multiple concurrent actions, should be used for benchmarking to better estimate spatio-temporal understanding.

6.2 Descriptive Video Understanding

Descriptive video understanding interprets video content through text descriptions. We compare state-of-the-art ViFMs on VideoQA and Video Captioning tasks within this category.

6.2.1 Video Question Answering (VideoQA)

Table 8 (left column) compares mean accuracy scores of ViFMs on VideoQA tasks on MSR-VTT [93], LSMDC [94], and MSVD [95] datasets. For VideoQA, different models excel depending on the question type. LAVENDER [248] (Video-based ViFM) performs best on multiple-choice (MC) and fill-in-the-blank (FiB) questions on MSR-VTT [93] and LSMDC [94], while Video-LLaVa [304] (Large Multimodal Model) excels at open-ended questions on [93] and MSVD [95].

LAVENDER's [248] success with MC and FiB questions is attributed to its strong cross-modal understanding, achieved by encoding both vision and text within a single joint encoder. Its MLM pretraining objective also aligns well with the FiB format. Conversely, open-ended questions benefit from combining video models with advanced LLMs, enhancing natural language understanding and video-text interaction.

Discussion. Beyond question types (MC vs. open-ended), VideoQA also relies on reasoning capability, divided into Factoid reasoning [39], [95] and Inference reasoning [346], [347]. Current ViFMs mainly address Factoid questions, which involve retrieving factual information directly from videos. However, there is a gap in handling Inference-based questions, which require a deeper understanding of video

	Method	Arch. Type	VQA					Captioning		
			MSRVTT		LSMDC		MSVD	MSRVTT	MSVD	YC2
			MC	QA	MC	FiB	QA			
	Fine-Tuned									
Generalist	ClipBERT [217]	DE	88.2	37.4	-	-	-	-	-	-
	RTQ [181]	DE	-	42.1	-	-	-	69.3	123.4	-
	Singularity [215]	DE	92.1	43.5	-	-	-	-	-	-
	UMT-B [110]	DE	96.3	44.9	-	-	49.5	-	-	-
	UMT-L [110]	DE	97.3	47.1	-	-	55.2	-	-	-
	VideoCoca [133]	ED	-	46.3	-	-	56.9	73.2	-	128.0
	All-in-One [144]	JE	92.3	46.8	84.4	-	48.3	-	-	-
	ALPRO [114]	ME	-	42.1	-	-	46.3	-	-	-
	Clover [111]	ME	95.2	44.1	83.7	54.1	52.4	-	-	-
	Hitea [33]	ME	97.2	45.4	85.8	54.6	55.6	62.5	145.1	-
	LAVANDER [248]	JE	**97.4**	45.0	**87.0**	**57.1**	56.6	60.1	150.7	-
	SimVTP [262]	ED	93.6	44.7	83.7	-	48.9	-	-	-
	UniVL [120]	ME	-	-	-	-	-	-	-	127.0
	VindLU-L [265]	DE	95.5	44.6	-	-	-	-	-	-
	VIOLET [143]	DE	91.9	43.9	82.8	53.7	47.9	-	-	-
	mPLUG-2 [296]	ME	-	48.0	-	-	58.1	**80.3**	165.8	-
	MaMMUT [192]	DE	-	49.5	-	-	60.2	73.6	**195.6**	-
	MERLOT [126]	DE	90.9	43.9	82.8	53.7	47.9	-	-	-
	OmniVL [282]	DE	-	44.1	-	-	51.0	-	-	116.0
	Smaug [294]	ME	92.9	44.5	-	-	-	-	-	-
	X^2VLM [135]	ME	-	45.5	-	-	54.6	-	-	-
	VALOR [145]	MLE	-	49.2	-	-	60.0	74.0	178.5	-
	VAST [113]	ME	-	50.1	-	-	60.2	78.0	-	**198.8**
	VideoChat [274]	LMM	-	45.0	-	-	56.3	-	-	-
	Video-LLaMA [7]	LMM	-	29.6	-	-	51.6	-	-	-
	Video-LLaVA [304]	LMM	-	**59.2**	-	-	**70.7**	-	-	-
	Video-ChatGPT [170]	LMM	-	49.3	-	-	64.9	-	-	-
Specialist	Just-Ask [343]	DE	-	41.5	-	-	46.3	-	-	-
	CLIP4Caption [164]	DE	-	-	-	-	-	57.7	-	-
	SwinBERT [344]	DE	-	-	-	-	-	53.8	120.6	-
	MV-GPT [156]	DE	-	-	-	-	-	60.0	-	-
	Text-KG [345]	ME	-	-	-	-	-	60.8	-	133
	Zero-Shot									
Generalist	VideoCoca [133]	ED	-	-	-	-	-	27.1	-	34.3
	Hitea [33]	ME	-	21.7	-	-	37.4	-	-	-
	mPlug-2 [296]	ME	-	**43.8**	-	-	**55.3**	-	-	-
	Distill-VLM [160]	LMM	-	24.4	-	-	-	**48.2**	-	-
	PaML2-VAdapter [202]	LMM	-	19.6	-	-	40.5	47.7	-	-
	VideoPrism-B [300] (1B)	DE	-	28.5	-	-	39.5	40.3	-	52.3
	VideoPrism-B [300] (8B)	DE	-	32.0	-	-	47.1	38.5	-	**63.6**

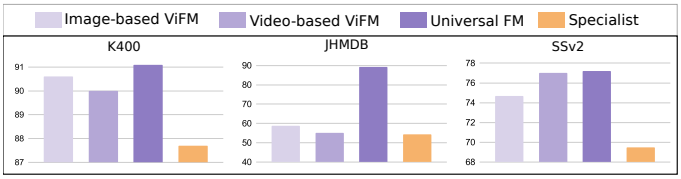
TABLE 8: Comparing the fine-tuned performance of SOTA generalist (Image-based, Video-based, UFMs) and specialist models for **VideoQA** (left) on MSRVTT [93], LSMDC [94], and MSVD [95] datasets; and fine-tuned, and Zero-shot **Video Captioning** (right)(CIDEr) tasks on MSRVTT [39], MSVD [99], and YC2/YouCook2 [41] datasets. The highlighted entries indicate the best performing methods.

content, including dense spatio-temporal relationships and causal relationships.

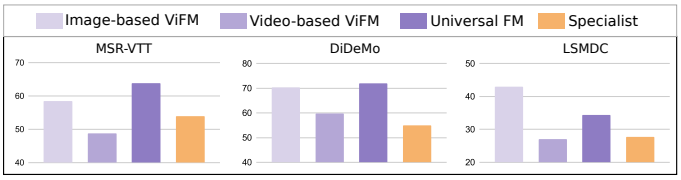
6.2.2 *Video Captioning.*

Table 8 (right column) compares CIDEr scores on MSRVTT [39], MSVD [99], and YouCook2 [41] datasets. The results show specialization: mPLUG-2 [296] excels on MSRVTT, MaMMUT [192] on MSVD, and VAST [113] on YouCook2 [41]. All these models are unified foundation models, highlighting the importance of relevant modalities for specific content. Notably, VAST's [113] strong performance on YouCook2 [41] underscores the value of audio in video captioning. Overall, unified models' superiority across datasets suggests the potential of incorporating additional modalities for captioning tasks.

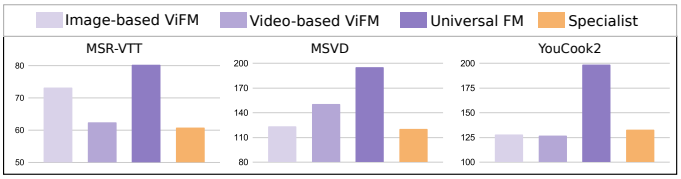
Discussion. While Table 8 focuses on coarse video captioning, dense video captioning, which generates multiple captions for events throughout a video, remains a challenge. This task requires a deeper understanding of spatio-temporal relationships, an area where ViFMs are still developing.



(a) Fine-tuning results for **video action recognition** task on k400 [23] (left), JHMDB [29] (center) and SSv2 [26] (right) datasets.



(b) Fine-tuning results for **video retrieval** task on MSRVTT [39] (left), DIDEMO [40] (center) and LSMDC [19] (right) datasets.



(c) Fine-tuning results for **video captioning** task on MSRVTT [39] (left), MSVD [99] (center) and YouCook2 [41] (right) datasets.

Fig. 6: Comparing fine-tuning results of Image-based, Video-based, and Universal foundational models against Specialist models. Best viewed in color.

6.3 Video Content Generation and Manipulation

Text-to-Video (T2V) generation creates videos from textual descriptions, primarily using GANs, auto-regressive, and diffusion models. In this section, we compare these models for the T2V generation task.

6.3.1 *Text-To-Video Generation*

Table 9 compares the performance of various models on the T2V generation task, using MSR-VTT [39] and UCF101 [30] datasets. For MSR-VTT [39], VideoPoet [13], employing a unified image-video framework, excels in both CLIPSim [106] and FVD [105] metrics. On UCF101, InternVid leads in the IS [107] metric, while SVD [315], a diffusion-based model, surpasses others in FVD, showcasing high visual fidelity.

	Method	MSR-VTT		UCF-101	
		ClipSIM (↑)	FVD (↓)	FVD (↓)	IS (↑)
Generalist	InternVid [162]	0.2951	-	616.51	21.04
	Make-A-Video [313]	**0.3049**	-	367.23	33.00
	PYoCo [314]	-	-	355.19	47.76
	SVD [315]	-	-	**242.02**	-
	VideoPoet [13]	**0.3049**	**213.00**	355.00	**38.44**
	VideoLaVIT [132]	0.3010	169.51	274.96	37.96
	VideoLDM [348]	0.2929	-	550.61	33.45
	VideoComposer [250]	0.2932	580.00	-	-
Special	CogVideo [309]	0.2631	1294.00	702	25.27
	MagicVideo [349]	-	998.00	655	-
	VideoFactory [350]	0.3005	-	410	-

TABLE 9: Comparing the fine-tuned performance of SOTA generalist and specialist models for zero-shot Text-To-Video Generation on MSR-VTT [39] and UCF-101 [30] dataset. The highlighted entries indicate the best performing methods.

Discussion. Despite advances in video generation, real-world applications remain challenging due to high computational demands and long processing times. Generating a minute of video can take hours. However, improvements in maintaining temporal consistency are promising, paving the way for integrating generative models with existing ViFMs to enhance representation spaces.

6.4 Analysis from Result Comparison

Figure 6 compares image-based ViFMs (light purple), video-based ViFMs (dark purple), and universal foundation models (UFMs) (darker purple) with task-specific models (orange) on video action recognition, retrieval, and captioning tasks. Task-specific models are fine-tuned for particular tasks.

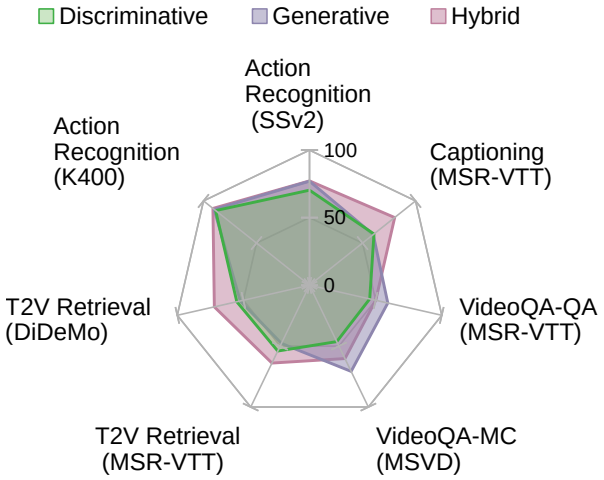


Fig. 7: Comparing ViFMs based on pretraining objectives (discriminative, generative, and hybrid). Best viewed in color.

A general trend is observed: image-based ViFMs outperform video-based ViFMs, likely due to the larger amount of pretraining data available for images compared to videos. An exception is video models excelling on SSv2 [26] in action recognition, likely due to the dataset's need for temporal understanding and sequence order. Universal models consistently outperform both image- and video-based ViFMs, indicating that integrating different modalities such as image,

video, and audio into a single framework provides significant performance gains. Notably, universal models surpass task-specific models for each case, demonstrating desirable generalization. This shows that training foundational models on massive datasets can outperform specialists and adapt to various downstream tasks with minimal compute.

We further compare ViFMs using discriminative, generative, and hybrid objectives during pretraining in Figure 7. This comparison reveals that foundation models with hybrid pretraining objectives generally perform better, except for the VQA task, where generative pretraining objectives (specifically LLM-based on next-token prediction) are more effective.

6.5 Summary

We compare Video Foundation Models (ViFMs) for video content understanding across three tasks, descriptive video understanding across two tasks, and video content generation on a single task. Some key observations among ViFMs and video modeling include the following: Mask Data Modeling (MDM), which also falls into the category of generative objectives, is a common pretraining objective for improving spatio-temporal modeling. The lack of availability and low performance of ViFMs on video action localization (see Table 7) indicate significant room for improvement in modeling spatio-temporal information. Revisiting spatio-temporal understanding from a research perspective is necessary.

Multimodal approaches, which usually employ vision-text contrastive (VTC) as a discriminative objective to improve semantic understanding by incorporating text with the visual modality, improve robustness and extend capabilities towards zero-shot [351] and open-vocabulary [352] understanding. However, ViFMs are far from achieving their full potential in these areas (see Tables 5, 6, and 8). A deeper analysis reveals a more critical issue: the limited scale of video-text datasets compared to image-text datasets. This data scarcity hinders effective learning for video foundation models. Additionally, the data quality of many video-text datasets, often generated by scraping online videos and text, introduces significant noise, further complicating the already challenging task of multi-modal learning for video models. Combining discriminative and generative objectives generally improves performance. However, a naive combination can degrade representations. Models like InterVideo2 [299] propose progressive training to address this issue, but further research is needed to optimize these hybrid objectives.

7 Challenges and Future Directions

Multi-modal video foundation models face significant challenges, including limited large-scale training data and high computational costs, impeding their development compared to image-text models. Existing video-text datasets are often small and noisy, hindering robust representation learning. Future research should prioritize creating high-quality, large-scale video-text datasets and exploring data cleaning, augmentation, and alternative pretraining methods.

Additionally, current models struggle with temporal consistency, object-centric tasks, and adaptability to diverse contexts. Recent advancements like the generative auto-regressive model SORA [311] offer valuable insights for temporal modeling. Extending ViFMs for fine-grained tasks requires careful consideration during pretraining [270].

With foundational models tackling an increasing number of tasks, comparing their performance using individual metrics is challenging. A shared metric using heuristics [279] can help, but aspects such as latency, energy consumption, memory usage, robustness, and task correlations must be considered. The rest of this section outlines potential future research directions for ViFMs.

Addressing Ethical Considerations. As Video Foundation Models (ViFMs) find increasing application in real-world scenarios, akin to ChatGPT [353] and Amazon SageMaker [354], addressing ethical concerns becomes crucial. Future research should focus on mitigating biases through debiasing datasets and fairness metrics, promoting transparency and explainability to build trust, and establishing responsible use guidelines throughout the ViFM development lifecycle. By actively addressing these ethical considerations, we can ensure ViFMs are deployed responsibly, maximizing their positive impact on real-world applications.

Long-Form Video Understanding. Achieving long-form video understanding [355] with ViFMs presents a significant challenge due to the extensive memory requirements for processing extended sequences. Recent efforts have recognized this hurdle and begun to explore solutions, such as memory consolidation mechanisms [277], [356] and memory-efficient attention [357]. However, to truly unlock the potential of ViFMs in this domain, integrating causal reasoning could be pivotal. By incorporating causal reasoning [347] into ViFMs, we can enhance their ability to comprehend extended video content by enabling them to answer fundamental questions like "why," "what next," and "what if." This deeper understanding facilitated by causal reasoning could revolutionize long-form video understanding, allowing ViFMs to recognize event order, direction of causality, and detailed relationships between actors, actions, and objects. Moreover, integrating causal reasoning may enhance the robustness [358] of ViFMs and improve their ability to handle occlusions [359] and other challenges commonly encountered in real-world video data. Therefore, while addressing the memory constraints is crucial, integrating causal reasoning into ViFMs offers a promising avenue for achieving comprehensive long-form video understanding.

Viewpoint Invariance. ViFMs excel in traditional video settings (i.e., third-person viewpoint), but limitations arise in understanding different viewpoints like egocentric [104] or birds-eye view [360]. Future research can delve into viewpoint-invariant representations for dynamic scenes. Inspired by the human ability to mentally rotate objects, neural representations like NeRF [361] can be explored to encode a continuous 3D representation within ViFMs. Additionally, methods that project an agent's limited view to a common reference frame [362] and establish correspondences across views hold promise for learning robust representations despite dynamic exploration. By pursuing these directions, ViFMs can be equipped to handle different viewpoints and varying camera paths, ultimately leading to significant advancements in action localization and a deeper understanding of dynamic video content.

Domain Adaptation Domain adaptation refers to the ability of models to perform well in new environments (lighting,

locations, etc.) or domains that differ from those they were trained on. Some studies [22], [363] provide directions to make traditional video models robust against such changes by suggesting augmentations [364] and specific tuning [365]. However, research in domain adaptation for foundation models is still lacking. To seamlessly integrate these powerful models into real-world applications, future work should explore methods to make ViFMs robust against domain shifts.

Improving Efficiency Despite promising results, ViFM's resource demands pose a challenge for edge deployment. These models often have hundreds of millions to billions of parameters, leading to longer training and inference times. This consequently limits their deployment on edge devices for real-time inference. To address this challenge, a key future direction in the integration of ViFMs for edge devices involves developing efficient deployment strategies to overcome resource constraints and enable seamless inferencing, e.g., VILA [366]. This entails exploring novel approaches to optimize model architecture and reduce computational overhead, as well as investigating innovative techniques for model compression and quantization to facilitate deployment on resource-constrained edge devices without compromising performance. By addressing these challenges, researchers can pave the way for widespread adoption of video foundation models in edge computing environments, unlocking their potential to power a diverse range of high-impact applications across industries.

8 Conclusion

This survey offers a comprehensive and, to the best of our knowledge, the first-of-its-kind in-depth exploration of Video Foundation Models (ViFMs). We commenced by establishing a foundation with discussions on video understanding tasks, relevant architectures, pretraining datasets, and approaches for ViFM pretraining. We categorize core methodologies for ViFM creation into three primary techniques: Image-based Models (inflating image foundation models for videos), Video-based models (focusing on video or video-text pre-training), and Universal Foundational Models (integrating multiple modalities apart from vision-text into the framework). By comparing the performance of various ViFMs on video tasks and offering insights based on methodologies and results, we aim to equip the research community with a comprehensive overview of existing ViFMs, while also highlighting critical areas for future exploration. This, we believe, will foster further advancements in video modeling and unlock the full potential of ViFMs.

9 Acknowledgements

We gratefully acknowledge the support of the Innovation Funds, Denmark, for funding this project. Grant number: 2081-00001B.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv*, 2021. 1

[2] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundational models defining a new era in vision: A survey and outlook," *arXiv:2307.13721*, 2023. 1

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of ICML*, pp. 8748–8763, 2021. 1, 5, 7, 8, 9, 10, 11, 12, 14

[4] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of ICML*, pp. 4904–4916, 2021. 1

[5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of ICCV*, pp. 4015–4026, 2023. 1, 10

[6] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," in *Proceedings of EMNLP*, pp. 6787–6800, 2021. 1, 5, 11, 12, 16

[7] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv*, 2023. 1, 13, 17

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of CVPR*, pp. 1725–1732, 2014. 1

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of CVPR*, pp. 6299–6308, 2017. 1

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of ICCV*, pp. 4489–4497, 2015. 1

[11] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of ICML*, 2021. 1, 11, 14

[12] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proceedings of NeurIPS*, p. 10078–10093, 2022. 1, 11, 15, 17

[13] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung, H. Adam, H. Akbari, Y. Alon, V. Birodkar, *et al.*, "Videopoet: A large language model for zero-shot video generation," *arXiv:2312.14125*, 2023. 1, 4, 5, 6, 13, 14, 18

[14] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of CVPR*, pp. 1933–1941, 2016. 1

[15] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of CVPR*, pp. 4584–4593, 2016. 1

[16] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang, "End-to-end learning of motion representation for video understanding," in *Proceedings of CVPR*, pp. 6016–6025, 2018. 1

[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of ICCV*, pp. 6836–6846, 2021. 1, 14

[18] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of ICCV*, pp. 4193–4202, 2017. 1

[19] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of CVPR*, pp. 3202–3212, 2015. 1, 2, 16, 17, 18

[20] T. Zhou, F. Porikli, D. J. Crandall, L. Van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE TPAMI*, vol. 45, no. 6, pp. 7099–7122, 2022. 1

[21] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Temporal sentence grounding in videos: A survey and future directions," *IEEE TPAMI*, 2023. 1

[22] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–37, 2023. 1, 20

[23] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 2, 7, 11, 14, 15, 16, 18

[24] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018. 2, 4, 7, 11

[25] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019. 2, 7, 8, 11, 14

[26] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the ICCV*, pp. 5842–5850, 2017. 2, 5, 7, 11, 14, 15, 16, 18

[27] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of CVPR*, pp. 961–970, 2015. 2, 3, 17

[28] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *Proceedings of ICCV*, pp. 8668–8678, 2019. 2, 3

[29] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Proceedings of ICCV*, pp. 2556–2563, 2011. 2, 11, 15, 16, 17, 18

[30] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012. 2, 3, 5, 11, 15, 16, 17, 18

[31] U. Demir, Y. S. Rawat, and M. Shah, "Tinyvirat: Low-resolution video action recognition," in *Proceedings of ICPR*, pp. 7387–7394, 2021. 2, 17

[32] Y. Li, Y. Li, and N. Vasconcelos, "Resound: Towards action recognition without representation bias," in *Proceedings of ECCV*, pp. 513–528, 2018. 2

[33] Q. Ye, G. Xu, M. Yan, H. Xu, Q. Qian, J. Zhang, and F. Huang, "Hitea: Hierarchical temporal-aware video-language pre-training," in *Proceedings of CVPR*, p. 15405–15416, 2023. 2, 4, 6, 11, 12, 16, 17

[34] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of ECCV*, pp. 720–736, 2018. 2, 4, 7, 11, 16

[35] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of ICCV*, pp. 706–715, 2017. 2

[36] A.-M. Oncescu, J. F. Henriques, Y. Liu, A. Zisserman, and S. Albanie, "Queryd: A video dataset with high-quality text and audio narrations," in *Proceedings of ICASSP*, pp. 2265–2269, 2021. 2

[37] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," in *Proceedings of ACCV*, 2020. 2

[38] Y. Wang, D. Gao, L. Yu, W. Lei, M. Feiszli, and M. Z. Shou, "Geb+: A benchmark for generic event boundary captioning, grounding and retrieval," in *Proceedings of ECCV*, pp. 709–725, 2022. 2

[39] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of CVPR*, pp. 5288–5296, 2016. 2, 4, 5, 16, 17, 18

[40] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of ICCV*, pp. 5803–5812, 2017. 2, 16, 18

[41] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of AAAI*, 2018. 2, 4, 17, 18

[42] K. Ashutosh, R. Girdhar, L. Torresani, and K. Grauman, "Hiervl: Learning hierarchical video-language embeddings," in *Proceedings of CVPR*, p. 23066–23078, 2023. 3, 4, 11, 12

[43] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision*, pp. 1–23, 2022. 3

[44] C. Lu, J. Shi, and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," in *Proceedings of ICCV*, pp. 2720–2727, 2013. 3

[45] W. Liu, W. Luo, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection – A New Baseline," in *Proceedings of CVPR*, pp. 6536–6545, 2018. 3

[46] B. Ramachandra and M. Jones, "Street Scene: A new dataset and evaluation protocol for video anomaly detection," in *Proceedings of WACV*, pp. 2569–2578, 2020. 3

[47] A. Acsintoae, A. Florescu, M. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "Ubnormal: New benchmark for supervised open-set video anomaly detection," in *Proceedings of CVPR*, pp. 20143–20153, 2022. 3

[48] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *Proceedings of CVPR*, pp. 6479–6488, 2018. 3

[49] Y. Sun, H. Xue, R. Song, B. Liu, H. Yang, and J. Fu, "Long-form video-language pre-training with multimodal temporal contrastive learning," in *Proceedings of NeurIPS*, p. 38032–38045, 2022. 3, 4, 5, 7, 11, 12, 16

[50] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of CVPR*, pp. 1207–1216, 2019. 3

[51] C.-Y. Wu and P. Krahenbuhl, "Towards long-form video understanding," in *Proceedings of CVPR*, pp. 1884–1894, 2021. 3

[52] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, S. Xing, G. Chen, J. Pan, J. Yu, Y. Wang, L. Wang, and Y. Qiao, "Internvideo: General video foundation models via generative and discriminative learning," *arXiv*, 2022. 3, 8, 13, 14, 17

[53] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *CVIU*, vol. 155, pp. 1–23, 2017. 3, 17

[54] Y. Liu, L. Wang, Y. Wang, X. Ma, and Y. Qiao, "Fineaction: A fine-grained video dataset for temporal action localization," *IEEE TIP*, vol. 31, pp. 6937–6950, 2022. 3

[55] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of CVPR*, pp. 780–787, 2014. 3

[56] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proceedings of ECCV*, pp. 510–526, 2016. 3

[57] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould, "The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose," in *Proceedings of WACV*, pp. 847–859, 2021. 3

[58] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of ICCV*, pp. 3192–3199, 2013. 3

[59] R. Modi, A. J. Rana, A. Kumar, P. Tirupattur, S. Vyas, Y. S. Rawat, and M. Shah, "Video action detection: Analysing limitations and challenges," *arXiv:2204.07892*, 2022. 3, 17

[60] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of CVPR*, pp. 6047–6056, 2018. 3, 14, 17

[61] A. Li, M. Thotakuri, D. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The ava-kinetics localized human actions video dataset (2020)," *arXiv:2005.00214*, 2005. 3, 17

[62] K. Duarte, Y. Rawat, and M. Shah, "Videocapsulenet: A simplified network for action detection," in *Proceedings of NeurIPS*, vol. 31, 2018. 3

[63] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, vol. 32, pp. 1231 – 1237, 2013. 3

[64] S. Rujikietgumjorn and N. Watcharapinchai, "Vehicle detection with sub-class training using r-cnn for the ua-detrac benchmark," *Proceedings of AVSS*, pp. 1–5, 2017. 3

[65] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," in *Proceedings of CVPR*, pp. 5374–5383, 2019. 3

[66] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv*, 2016. 3

[67] P. Dendorfer, H. Rezatofighi, A. Milan, J. Q. Shi, D. Cremers, I. D. Reid, S. Roth, K. Schindler, and L. Leal-Taix'e, "Mot20: A benchmark for multi object tracking in crowded scenes," *ArXiv*, 2020. 3

[68] M. Fabbri, G. Brasó, G. Maugeri, O. Cetintas, R. Gasparini, A. Ošep, S. Calderara, L. Leal-Taixé, and R. Cucchiara, "Motsynth: How can synthetic data help pedestrian detection and tracking?," in *Proceedings of ICCV*, pp. 10849–10859, 2021. 3

[69] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of CVPR*, pp. 2636–2645, 2020. 3, 7, 14

[70] A. Dave, T. Khurana, P. Tokmakov, C. Schmid, and D. Ramanan, "Tao: A large-scale benchmark for tracking any object," in *Proceedings of ECCV*, pp. 436–454, 2020. 3

[71] A. Athar, J. Luiten, P. Voigtlaender, T. Khurana, A. Dave, B. Leibe, and D. Ramanan, "Burst: A benchmark for unifying object recognition, segmentation and tracking in video," in *Proceedings of WACV*, pp. 1674–1683, 2023. 3

[72] H. Wang, C. Yan, S. Wang, X. Jiang, X. Tang, Y. Hu, W. Xie, and E. Gavves, "Towards open-vocabulary video instance segmentation," in *Proceedings of ICCV*, pp. 4057–4066, 2023. 3

[73] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *IJCV*, vol. 129, pp. 548–578, 2021. 3

[74] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008. 3

[75] Y. Zheng, A. W. Harley, B. Shen, G. Wetzstein, and L. J. Guibas, "Pointodyssey: A large-scale synthetic dataset for long-term point tracking," in *Proceedings of ICCV*, pp. 19855–19865, 2023. 3

[76] C. Doersch, A. Gupta, L. Markeeva, A. Recasens, L. Smaira, Y. Aytar, J. Carreira, A. Zisserman, and Y. Yang, "Tap-vid: A benchmark for tracking any point in a video," in *Proceedings of NeurIPS*, pp. 13610–13626, 2022. 3

[77] R. Sundararaman, C. De Almeida Braga, E. Marchand, and J. Pettre, "Tracking pedestrian heads in dense crowd," in *Proceedings of CVPR*, pp. 3865–3875, 2021. 3

[78] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang, "Youtube-vos: A large-scale video object segmentation benchmark," *arXiv:1809.03327*, 2018. 3

[79] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proceedings of ICCV*, pp. 5188–5197, 2019. 3, 7, 14

[80] X. Ning, Y. Linjie, Y. Jianchao, Y. Dingcheng, F. Yuchen, L. Yuchen, and S. H. Thomas, "Youtubevis dataset 2021 version." https://youtube-vos.org/dataset/vis/, 2021. 3, 7, 14

[81] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of CVPR*, pp. 724–732, 2016. 3, 5

[82] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Video panoptic segmentation," in *Proceedings of CVPR*, pp. 9859–9868, 2020. 3

[83] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Proceedings of ECCV*, pp. 792–807, 2016. 3, 7, 14

[84] S. Seo, J.-Y. Lee, and B. Han, "Urvos: Unified referring video object segmentation network with a large-scale benchmark," in *Proceeding of the ECCV*, pp. 208–223, 2020. 3, 14

[85] A. Khoreva, A. Rohrbach, and B. Schiele, "Video object segmentation with language referring expressions," in *Proceedings of ACCV*, pp. 123–141, 2019. 3

[86] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek, "Actor and action video segmentation from a sentence," in *Proceedings of CVPR*, pp. 5958–5966, 2018. 3

[87] K. Duarte, Y. S. Rawat, and M. Shah, "Capsulevos: Semi-supervised video object segmentation using capsule routing," in *Proceedings of ICCV*, pp. 8480–8489, 2019. 3

[88] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015. 4

[89] Z. Zhang, Z. Zhao, Y. Zhao, Q. Wang, H. Liu, and L. Gao, "Where does it exist: Spatio-temporal video grounding for multi-form sentences," in *Proceedings of CVPR*, pp. 10668–10677, 2020. 4

[90] Z. Tang, Y. Liao, S. Liu, G. Li, X. Jin, H. Jiang, Q. Yu, and D. Xu, "Human-centric spatio-temporal video grounding with visual transformers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 8238–8249, 2020. 4

[91] M. Mathew, D. Karatzas, and C. Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of WACV*, pp. 2200–2209, 2021. 4

[92] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of CVPR*, pp. 2758–2766, 2017. 4

[93] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proceedings of ECCV*, pp. 471–487, 2018. 4, 17

[94] A. Torabi, N. Tandon, and L. Sigal, "Learning language-visual embedding for movie understanding with natural-language," *arXiv:1609.08124*, 2016. 4, 17

[95] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of ACM-MM*, pp. 1645–1653, 2017. 4, 17

[96] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of AAAI*, pp. 9127–9134, 2019. 4

[97] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+ language omni-representation pre-training," *arXiv:2005.00200*, 2020. 4

[98] J. Liu, W. Chen, Y. Cheng, Z. Gan, L. Yu, Y. Yang, and J. Liu, "Violin: A large-scale dataset for video-and-language inference. in 2020 ieee," in *Proceedings of CVPR*, pp. 13–19, 2020. 4

[99] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of ACL:HLT*, pp. 190–200, 2011. 4, 17, 18

[100] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, pp. 311–318, 2002. 4

[101] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL-W*, pp. 65–72, 2005. 4

[102] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of ACL-W*, pp. 74–81, 2004. 4

[103] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of CVPR*, pp. 4566–4575, 2015. 4

[104] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of CVPR*, pp. 18995–19012, 2022. 4, 11, 16, 19

[105] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv:1812.01717*, 2018. 5, 18

[106] C. Wu, L. Huang, Q. Zhang, B. Li, L. Ji, F. Yang, G. Sapiro, and N. Duan, "Godiva: Generating open-domain videos from natural descriptions," *arXiv:2104.14806*, 2021. 5, 18

[107] M. Saito, S. Saito, M. Koyama, and S. Kobayashi, "Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan," *IJCV*, vol. 128, no. 10-11, pp. 2586–2606, 2020. 5, 18

[108] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of CVPR*, p. 14549–14560, 2023. 5, 6, 11, 15, 17

[109] C. Feichtenhofer, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," in *Proceedings of NeurIPS*, vol. 35, p. 35946–35958, 2022. 5, 11, 15, 17

[110] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao, "Unmasked teacher: Towards training-efficient video foundation models," in *Proceedings of ICCV*, 2023. 5, 7, 8, 12, 15, 16, 17

[111] J. Huang, Y. Li, J. Feng, X. Wu, X. Sun, and R. Ji, "Clover: Towards a unified video-language alignment and fusion model," in *Proceedings of CVPR*, pp. 14856–14866, 2023. 5, 6, 11, 12, 16, 17

[112] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proceedings of CVPR*, p. 16102–16112, 2022. 5, 13

[113] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu, "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset," *arXiv*, 2023. 5, 6, 7, 14, 15, 16, 17, 18

[114] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of CVPR*, p. 4953–4963, 2022. 5, 6, 7, 11, 12, 16, 17

[115] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "Videomoco: Contrastive video representation learning with temporally adversarial examples," in *Proceedings of CVPR*, pp. 11205–11214, 2021. 5

[116] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent – a new approach to self-supervised learning," in *Proceedings of NeurIPS*, pp. 21271–21284, 2020. 5

[117] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "Simclr: A simple framework for contrastive learning of visual representations," in *Proceedings of ICLR*, 2020. 5, 11

[118] H. Tan, J. Lei, T. Wolf, and M. Bansal, "Vimpac: Video pre-training via masked token prediction and contrastive learning," *arXiv preprint arXiv:2106.11250*, 2021. 5, 10, 11, 12, 15

[119] L. Momeni, M. Caron, A. Nagrani, A. Zisserman, and C. Schmid, "Verbs in action: Improving verb understanding in video-language models," in *Proceedings of ICCV*, p. 15579–15591, 2023. 5, 11, 12

[120] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv*, 2020. 5, 6, 11, 12, 16, 17, 29

[121] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass, "Contrastive audio-visual masked autoencoder," in *Proceedings of ICLR*, 2023. 6, 14, 15

[122] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Proceedings of NeurIPS*, p. 24206–24221, 2021. 6, 14, 15

[123] N. Parthasarathy, S. A. Eslami, J. Carreira, and O. J. Henaff, "Self-supervised video pretraining yields robust and more human-aligned visual representations," in *Proceedings of NeurIPS*, 2023. 6, 14

[124] P. Bagad, M. Tapaswi, and C. G. Snoek, "Test of time: Instilling video-language models with a sense of time," in *Proceedings of CVPR*, pp. 2503–2516, 2023. 6, 11, 12

[125] Z. Wang, A. Blume, S. Li, G. Liu, J. Cho, Z. Tang, M. Bansal, and H. Ji, "Paxion: Patching action knowledge in video-language foundation models," in *Proceedings of NeurIPS*, 2023. 6, 11, 12

[126] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," in *Proceedings of NeurIPS*, pp. 23634–23651, 2021. 6, 14, 17

[127] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *Proceedings of CVPR*, p. 5036–5045, 2022. 6, 7, 8, 11, 12, 14, 16

[128] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "An empirical study of end-to-end video-language transformers with masked visual modeling," in *Proceedings of CVPR*, p. 22898–22909, 2023. 6, 7, 11, 12

[129] B. Huang, Z. Zhao, G. Zhang, Y. Qiao, and L. Wang, "Mgmae: Motion guided masking for video masked autoencoding," in *Proceedings of ICCV*, p. 13493–13504, 2023. 6, 11, 15

[130] H. Xu, G. Ghosh, P.-Y. Huang, P. Arora, M. Aminzadeh, C. Feichtenhofer, F. Metze, and L. Zettlemoyer, "Vlm: Task-agnostic video-language model pre-training for video understanding," in *Proceedings of ACL-IJCNLP*, pp. 4227–4239, 2021. 6, 13, 14

[131] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, and L. Yuan, "Bevt: Bert pretraining of video transformers," in *Proceedings of CVPR*, p. 14733–14743, 2022. 6, 11, 15

[132] Y. Jin, Z. Sun, K. Xu, L. Chen, H. Jiang, Q. Huang, C. Song, Y. Liu, D. Zhang, Y. Song, *et al.*, "Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization," *arXiv:2402.03161*, 2024. 6, 13, 14, 18

[133] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, "Videococa: Video-text modeling with zero-shot transfer from contrastive captioners," *arXiv*, 2023. 6, 7, 8, 15, 16, 17

[134] T.-J. Fu, L. Yu, N. Zhang, C.-Y. Fu, J.-C. Su, W. Y. Wang, and S. Bell, "Tell me what happened: Unifying text-guided video completion via multimodal masked video generation," in *Proceedings of CVPR*, p. 10681–10692, 2023. 6, 11, 12

[135] Y. Zeng, X. Zhang, H. Li, J. Wang, J. Zhang, and W. Zhou, "X^2-vlm: All-in-one pre-trained model for vision-language tasks," *IEEE TPAMI*, vol. 46, no. 5, pp. 3156–3168, 2024. 6, 14, 15, 16, 17

[136] J. Xu, B. Liu, Y. Chen, M. Cheng, and X. Shi, "Multi: Efficient video-and-language understanding with multiway-sampler and multiple choice modeling," *arXiv*, 2023. 6, 11, 12, 13

[137] S. Castro and F. C. Heilbron, "Fitclip: Refining large-scale pre-trained image-text models for zero-shot video understanding tasks," in *Proceedings of BMVC*, 2022. 6, 7, 8, 16

[138] B. Zhang, Y. Ge, X. Xu, Y. Shan, and M. Z. Shou, "Taca: Upgrading your visual foundation model with task-agnostic compatible adapter," *arXiv*, 2023. 6, 7, 8, 9

[139] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of CVPR*, pp. 1728–1738, 2021. 7, 8, 11, 14, 15, 16

[140] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of ICCV*, pp. 2630–2640, 2019. 7, 8, 11, 14

[141] T. Gupta, D. Schwenk, A. Farhadi, D. Hoiem, and A. Kembhavi, "Imagine this! scripts to compositions to videos," in *Proceedings of ECCV*, pp. 598–613, 2018. 7, 11

[142] T. Hayes, S. Zhang, X. Yin, G. Pang, S. Sheng, H. Yang, S. Ge, Q. Hu, and D. Parikh, "Mugen: A playground for video-audio-text multimodal understanding and generation," in *Proceedings of ECCV*, pp. 431–449, 2022. 7, 11

[143] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "Violet: End-to-end video-language transformers with masked visual-token modeling," *arXiv*, 2022. 7, 11, 12, 16, 17

[144] J. Wang, Y. Ge, R. Yan, Y. Ge, K. Q. Lin, S. Tsutsui, X. Lin, G. Cai, J. Wu, and Y. Shan, "All in one: Exploring unified video-language pre-training," in *Proceedings of CVPR*, p. 6598–6608, 2023. 7, 11, 12, 15, 16, 17

[145] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *arXiv*, 2023. 7, 8, 14, 15, 16, 17

[146] Z. Tang, J. Cho, J. Lei, and M. Bansal, "Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention," in *Proceedings of WACV*, p. 4410–4420, 2023. 7, 14, 15

[147] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of ACL*, pp. 2556–2565, 2018. 7, 8, 11, 14

[148] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of CVPR*, pp. 3558–3568, 2021. 7, 8, 11, 14

[149] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proceedings of NeurIPS*, 2011. 7, 8, 14

[150] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *Proceedings of ICCV*, pp. 8430–8439, 2019. 7, 14

[151] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, vol. 128, no. 7, pp. 1956–1981, 2020. 7, 14

[152] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of ECCV*, pp. 740–755, 2014. 7, 8, 14

[153] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proceedings of ECCV*, pp. 69–85, 2016. 7, 14

[154] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, pp. 32–73, 2017. 7, 8, 14

[155] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of CVPR*, pp. 5356–5364, 2019. 7, 14

[156] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proceedings of CVPR*, p. 17959–17968, 2022. 7, 17

[157] J. Qi, Y. Gao, Y. Hu, X. Wang, X. Liu, X. Bai, S. Belongie, A. Yuille, P. H. Torr, and S. Bai, "Occluded video instance segmentation: A benchmark," *IJCV*, vol. 130, no. 8, pp. 2022–2039, 2022. 7, 14

[158] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv:2306.14824*, 2023. 7, 14

[159] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of ICCV*, pp. 4015–4026, 2023. 7, 10, 14

[160] Y. Zhao, L. Zhao, X. Zhou, J. Wu, C.-T. Chu, H. Miao, F. Schroff, H. Adam, T. Liu, B. Gong, *et al.*, "Distilling vision-language models on millions of videos," *arXiv preprint arXiv:2401.06129*, 2024. 7, 8, 17

[161] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid, "Learning audio-video modalities from image captions," in *Proceedings of ECCV*, pp. 407–426, 2022. 7, 8

[162] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," in *Proceedings of ICLR*, 2023. 7, 11, 12, 17, 18

[163] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, p. 293–304, 2022. 7, 8, 16

[164] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, and X. Li, "Clip4caption: Clip for video caption," in *Proceedings of ACM-MM*, pp. 4858–4862, 2021. 7, 8, 17

[165] Y. Li, K. Li, Y. He, Y. Wang, Y. Wang, L. Wang, Y. Qiao, and P. Luo, "Harvest video foundation models via efficient post-pretraining," *arXiv*, 2023. 7, 8

[166] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of NeurIPS*, 2023. 7

[167] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of ICML*, pp. 12888–12900, 2022. 7, 8, 14

[168] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023. 7, 10

[169] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023. 7, 8

[170] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv*, 2023. 7, 11, 14, 17

[171] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *arXiv*, 2023. 7

[172] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *Proceedings of ICML*, pp. 2790–2799, 2019. 7

[173] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2video: Mastering video-text retrieval via image clip," *arXiv:2106.11097*, 2021. 7

[174] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, and M. Li, "Aim: Adapting image models for efficient video action recognition," in *Proceedings of ICLR*, 2022. 7, 8, 9, 15, 16

[175] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "St-adapter: Parameter-efficient image-to-video transfer learning," *Proceedings of NeurIPS*, pp. 26462–26477, 2022. 7, 9, 15, 16

[176] M. Wang, J. Xing, B. Jiang, J. Chen, J. Mei, X. Zuo, G. Dai, J. Wang, and Y. Liu, "M2-clip: A multimodal, multi-task adapting framework for video action recognition," *arXiv:2401.11649*, 2024. 7, 9, 15, 16

[177] X. Li and L. Wang, "Zeroi2v: Zero-cost adaptation of pre-trained transformers from image to video," *arXiv*, 2023. 7, 9

[178] H. Fang, Z. Yang, Y. Wei, X. Zang, C. Ban, Z. Feng, Z. He, Y. Li, and H. Sun, "Alignment and generation adapter for efficient video-text understanding," in *Proceedings of ICCV*, p. 2791–2797, 2023. 7, 8, 9

[179] Z. Lin, S. Geng, R. Zhang, P. Gao, G. De Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," in *Proceedings of ECCV*, p. 388–404, 2022. 7, 8, 9, 15

[180] Z. Qing, S. Zhang, Z. Huang, Y. Zhang, C. Gao, D. Zhao, and N. Sang, "Disentangling spatial and temporal learning for efficient image-to-video transfer learning," in *Proceedings of ICCV*, p. 13934–13944, 2023. 7, 8, 9

[181] X. Wang, Y. Li, T. Gan, Z. Zhang, J. Lv, and L. Nie, "Rtq: Rethinking video-language understanding based on image-text model," in *Proceedings of ACM-MM*, p. 557–566, 2023. 7, 8, 9, 16, 17

[182] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of EMNLP*, pp. 3045–3059, 2021. 7

[183] S. Huang, B. Gong, Y. Pan, J. Jiang, Y. Lv, Y. Li, and D. Wang, "Vop: Text-video co-operative prompt tuning for cross-modal retrieval," in *Proceedings of CVPR*, p. 6565–6574, 2023. 7, 9, 11

[184] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *Proceedings of CVPR*, p. 23034–23044, 2023. 7, 9

[185] D. Engin and Y. Avrithis, "Zero-shot and few-shot video question answering with multi-modal prompts," in *Proceedings of ICCV*, p. 2804–2810, 2023. 7, 10

[186] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *Proceedings of ECCV*, pp. 105–124, 2022. 7, 8, 10

[187] A. Yousaf, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Video-prompter: an ensemble of foundational models for zero-shot video understanding," *arXiv preprint arXiv:2310.15324*, 2023. 7, 8, 9

[188] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 8, 11, 14

[189] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of NeurIPS*, 2017. 8, 11, 14

[190] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, "Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer," *arXiv preprint arXiv:2211.09552*, 2022. 8, 14

[191] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. 8, 11, 12, 14

[192] W. Kuo, A. Piergiovanni, D. Kim, B. Caine, W. Li, A. Ogale, L. Zhou, A. M. Dai, Z. Chen, C. Cui, *et al.*, "Mammut: A simple architecture for joint learning for multimodal tasks," *Transactions on Machine Learning Research*, 2023. 8, 17, 18

[193] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceeedings of ICML*, pp. 4904–4916, 2021. 8

[194] H. Xue, Y. Sun, B. Liu, J. Fu, R. Song, H. Li, and J. Luo, "Clip-vip: Adapting pre-trained image-text model to video-language representation alignment," *arXiv*, 2023. 8

[195] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, and A. Oliva, "Spoken moments: Learning joint audio-visual representations from video descriptions," in *Proceedings of CVPR*, pp. 14871–14881, 2021. 8

[196] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv:2209.06794*, 2022. 8

[197] X. Chen, X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, B. Mustafa, S. Goodman, I. Alabdulmohsin, P. Padlewski, *et al.*, "Pali-3 vision language models: Smaller, faster, stronger," *arXiv:2310.09199*, 2023. 8

[198] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of CVPR*, pp. 12104–12113, 2022. 8

[199] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, D. Bahri, T. Schuster, S. Zheng, *et al.*, "Ul2: Unifying language learning paradigms," in *Proceedings of ICLR*, 2022. 8

[200] J. Park, J. Lee, and K. Sohn, "Dual-path adaptation from image to video transformers," in *Proceedings of CVPR*, p. 2203–2213, 2023. 8, 9, 15

[201] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv:2111.02114*, 2021. 8, 11

[202] J. Xiao, Z. Xu, A. Yuille, S. Yan, and B. Wang, "Palm2-vadapter: Progressively aligned language model makes a strong vision-language adapter," *arXiv:2402.10896*, 2024. 8, 9, 12, 17

[203] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," in *Proceedings of NeurIPS*, pp. 23716–23736, 2022. 8

[204] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv:2205.01917*, 2022. 8

[205] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, *et al.*, "Palm 2 technical report," *arXiv:2305.10403*, 2023. 8

[206] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the" video" in video-language understanding," in *Proceedings of CVPR*, pp. 2917–2927, 2022. 8, 9, 10

[207] Y. Jian, C. Gao, and S. Vosoughi, "Bootstrapping vision-language learning with decoupled language pre-training," *arXiv:2307.07063*, 2023. 8

[208] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *Proceedings of NeurIPS*, vol. 35, pp. 25278–25294, 2022. 8

[209] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023. 8, 14, 16

[210] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, "Opt: Open pre-trained transformer language models," *arXiv:2205.01068*, 2022. 8

[211] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv:2301.12597*, 2023. 8

[212] OpenAI, "Gpt-4 technical report," *arXiv:2303.08774*, 2023. 8

[213] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-tuned clip models are efficient video learners," in *Proceedings of CVPR*, p. 6545–6554, 2023. 8, 10, 16

[214] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv:2109.08472*, 2021. 8, 15, 16

[215] J. Lei, T. L. Berg, and M. Bansal, "Revealing single frame bias for video-and-language learning," *arXiv*, 2022. 8, 10, 16, 17

[216] S. Chen, X. He, H. Li, X. Jin, J. Feng, and J. Liu, "Cosa: Concatenated sample pretrained vision-language foundation model," *arXiv*, 2023. 8, 10, 29

[217] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," in *Proceedings of CVPR*, pp. 7331–7341, 2021. 8, 17

[218] C. Deng, Q. Chen, P. Qin, D. Chen, and Q. Wu, "Prompt switch: Efficient clip adaptation for text-video retrieval," in *Proceedings of ICCV*, p. 15648–15658, 2023. 9

[219] H. Fang, P. Xiong, L. Xu, and W. Luo, "Transferring image-clip to video-text retrieval via temporal relations," *IEEE Transactions on Multimedia*, 2022. 9

[220] H. Jiang, J. Zhang, R. Huang, C. Ge, Z. Ni, J. Lu, J. Zhou, S. Song, and G. Huang, "Cross-modal adapter for text-video retrieval," *arXiv*, 2022. 9

[221] Z. Hu, A. N. Ye, S. Hosseini Khorasgani, and I. Mohomed, "Adaclip: Towards pragmatic multimodal video retrieval," in *Proceedings of ACM-MM*, p. 5623–5633, 2023. 9

[222] Z. Dai, F. Shao, Q. Su, Z. Dong, and S. Zhu, "Fine-grained text-video retrieval with frozen image encoders," *arXiv preprint arXiv:2307.09972*, 2023. 9, 16

[223] G. Chen, X. Liu, G. Wang, K. Zhang, P. H. Torr, X.-P. Zhang, and Y. Tang, "Tem-adapter: Adapting image-text pretraining for video question answer," in *Proceedings of ICCV*, p. 13945–13955, 2023. 9

[224] T. Phan, K. Vo, D. Le, G. Doretto, D. Adjeroh, and N. Le, "Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection," *arXiv*, 2023. 9

[225] Y. Jian, C. Gao, and S. Vosoughi, "Bootstrapping vision-language learning with decoupled language pre-training," *arXiv*, 2023. 9

[226] D. Romero and T. Solorio, "Question-instructed visual descriptions for zero-shot video question answering," *arXiv:2402.10698*, 2024. 10

[227] Q. Wang, J. Du, K. Yan, and S. Ding, "Seeing in flowing: Adapting clip for action recognition with motion prompts learning," in *Proceedings of ACM Multimedia*, pp. 5339–5347, 2023. 10

[228] S. Zhao, L. Zhu, X. Wang, and Y. Yang, "Centerclip: Token clustering for efficient text-video retrieval," in *Proceedings of SIGIR*, p. 970–981, 2022. 10

[229] S. K. Gorti, N. Vouitsis, J. Ma, K. Golestan, M. Volkovs, A. Garg, and G. Yu, "X-pool: Cross-modal language-video attention for text-video retrieval," in *Proceedings of CVPR*, p. 5006–5015, 2022. 10

[230] S.-M. Kang and Y.-S. Cho, "Meme: Multi-encoder multi-expert framework with data augmentation for video retrieval," in *Proceedings of SIGIR*, pp. 475–484, 2023. 10

[231] P. Li, C.-W. Xie, L. Zhao, H. Xie, J. Ge, Y. Zheng, D. Zhao, and Y. Zhang, "Progressive spatio-temporal prototype matching for text-video retrieval," in *Proceedings of ICCV*, pp. 4100–4110, 2023. 10, 16

[232] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang, "Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models," in *Proceedings of CVPR*, p. 6620–6630, 2023. 10, 15, 16

[233] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," *arXiv*, 2023. 10

[234] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," *arXiv*, 2023. 10

[235] J. Zhu, Z. Chen, Z. Hao, S. Chang, L. Zhang, D. Wang, H. Lu, B. Luo, J.-Y. He, J.-P. Lan, H. Chen, and C. Li, "Tracking anything in high quality," *arXiv*, 2023. 10

[236] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee, "Tracking anything with decoupled video segmentation," in *Proceedings of ICCV*, pp. 1316–1326, 2023. 10

[237] Z. Zhang, Z. Wei, S. Zhang, Z. Dai, and S. Zhu, "Uvosam: A mask-free paradigm for unsupervised video object segmentation via segment anything model," *arXiv*, 2023. 10

[238] Y. Li, J. Zhang, X. Teng, and L. Lan, "Refsam: Efficiently adapting segmenting anything model for referring video object segmentation," *arXiv*, 2023. 10

[239] F. Rajič, L. Ke, Y.-W. Tai, C.-K. Tang, M. Danelljan, and F. Yu, "Segment anything meets point tracking," *arXiv*, 2023. 10

[240] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of CVPR*, pp. 248–255, 2009. 11, 14

[241] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of CVPR*, pp. 6299–6308, 2017. 11, 14

[242] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of CVPR*, pp. 3202–3211, 2022. 11, 12, 14

[243] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," in *Proceedings of NeurIPS*, 2017. 11, 14

[244] Y. Song, M. Yang, W. Wu, D. He, F. Li, and J. Wang, "It takes two: Masked appearance-motion modeling for self-supervised video transformer pre-training," *arXiv*, 2022. 11, 15

[245] M.-I. Georgescu, E. Fonseca, R. T. Ionescu, M. Lucic, C. Schmid, and A. Arnab, "Audiovisual masked autoencoders," in *Proceedings of ICCV*, p. 16144–16154, 2023. 11, 12

[246] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *Proceedings of ICASSP*, pp. 721–725, IEEE, 2020. 11

[247] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of CVPR*, pp. 16000–16009, 2022. 10, 11, 14

[248] L. Li, Z. Gan, K. Lin, C.-C. Lin, Z. Liu, C. Liu, and L. Wang, "Lavender: Unifying video-language understanding as masked language modeling," in *Proceedings of CVPR*, p. 23119–23129, 2023. 11, 16, 17

[249] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, L. Yuan, and Y.-G. Jiang, "Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning," in *Proceedings of CVPR*, p. 6312–6322, 2023. 11

[250] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," in *Proceedings of NeurIPS*, 2024. 11, 13, 18

[251] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of NeurIPS*, pp. 6840–6851, 2020. 11, 14

[252] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of ICML*, pp. 2256–2265, 2015. 11, 14

[253] S. Hwang, J. Yoon, Y. Lee, and S. J. Hwang, "Efficient video representation learning via motion-aware token selection," *arXiv*, 2023. 11

[254] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of ICCV*, p. 7464–7473, 2019. 11, 12

[255] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019. 11

[256] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, and A. Oliva, "Spoken moments: Learning joint audio-visual representations from video descriptions," in *Proceedings of CVPR*, pp. 14871–14881, 2021. 11

[257] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *JMLR*, vol. 24, no. 240, pp. 1–113, 2023. 11

[258] Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar, "Learning video representations from large language models," in *Proceedings of CVPR*, p. 6586–6597, 2023. 11, 12

[259] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019. 11

[260] Z. Tang, J. Cho, Y. Nie, and M. Bansal, "Tvlt: Textless vision-language transformer," *Proceedings of NeurIPS*, p. 9617–9632, 2022. 11, 12

[261] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," in *Poceedings of NeurIPS*, pp. 23634–23651, 2021. 11, 14

[262] Y. Ma, T. Yang, Y. Shan, and X. Li, "Simvtp: Simple video text pre-training with masked autoencoders," *arXiv*, 2022. 11, 12, 13, 16, 17

[263] B. Zhang, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016. 11

[264] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of CVPR*, pp. 4804–4814, 2022. 11, 14

[265] F. Cheng, X. Wang, J. Lei, D. Crandall, M. Bansal, and G. Bertasius, "Vindlu: A recipe for effective video-and-language pretraining," in *Proceedings of CVPR*, pp. 10739–10750, 2023. 11, 13, 16, 17

[266] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of ICCV*, pp. 12179–12188, 2021. 11

[267] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proceedings of ECCV*, pp. 402–419, 2020. 11

[268] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of ICCV*, pp. 10012–10022, 2021. 11, 14

[269] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proceedings of ICML*, pp. 8821–8831, 2021. 11

[270] N. Madan, N.-C. Ristea, K. Nasrollahi, T. B. Moeslund, and R. T. Ionescu, "Cl-mae: Curriculum-learned masked autoencoders," in *Proceedings of WACV*, pp. 2492–2502, 2024. 11, 19

[271] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of CVPR*, pp. 12873–12883, 2021. 11

[272] J. Wang, D. Chen, C. Luo, X. Dai, L. Yuan, Z. Wu, and Y.-G. Jiang, "Chatvideo: A tracklet-centric multimodal and versatile video understanding system," *arXiv:2304.14407*, 2023. 11

[273] K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, and L. Wang, "Mm-vid: Advancing video understanding with gpt-4v(ision)," *arXiv*, 2023. 11

[274] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv*, 2023. 11, 17

[275] R. Luo, Z. Zhao, M. Yang, J. Dong, D. Li, P. Lu, T. Wang, L. Hu, M. Qiu, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," *arXiv*, 2023. 11

[276] H. Lin, A. Zala, J. Cho, and M. Bansal, "Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning," *arXiv*, 2023. 12, 13, 14

[277] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, X. Guo, T. Ye, Y. Lu, J.-N. Hwang, and G. Wang, "Moviechat: From dense token to sparse memory for long video understanding," *arXiv*, 2023. 12, 19

[278] D. Ko, J. Choi, H. K. Choi, K.-W. On, B. Roh, and H. J. Kim, "Meltr: Meta loss transformer for learning to fine-tune video foundation models," in *Proceedings of CVPR*, p. 20105–20115, 2023. 13

[279] L. Yuan, N. B. Gundavarapu, L. Zhao, H. Zhou, Y. Cui, L. Jiang, X. Yang, M. Jia, T. Weyand, L. Friedman, *et al.*, "Videoglue: Video general understanding evaluation of foundation models," *arXiv:2307.03166*, 2023. 13, 19

[280] M. Ning, B. Zhu, Y. Xie, B. Lin, J. Cui, L. Yuan, D. Chen, and L. Yuan, "Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models," *arXiv*, 2023. 13

[281] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proceedings of CVPR*, pp. 14668–14678, 2022. 13, 14, 15, 17

[282] J. Wang, D. Chen, Z. Wu, C. Luo, L. Zhou, Y. Zhao, Y. Xie, C. Liu, Y.-G. Jiang, and L. Yuan, "Omnivl: One foundation model for image-language and video-language tasks," in *Proceedings of NeurIPS*, p. 5696–5710, 2022. 13, 14, 15, 16, 17

[283] R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Omnimae: Single model masked pretraining on images and videos," in *Proceedings of CVPR*, p. 10406–10417, 2023. 13, 14, 15

[284] T. Computer, "Redpajama: an open dataset for training large language models," October 2023. 14

[285] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv:2310.03744*, 2023. 14

[286] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, D. Bahri, T. Schuster, S. Zheng, *et al.*, "Ul2: Unifying language learning paradigms," in *Proceedings of ICLR*, 2022. 14

[287] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of ICASSP*, pp. 776–780, 2017. 14

[288] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, W. Zhang, Z. Li, W. Liu, and L. Yuan, "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," in *Proceedings of ICLR*, 2024. 14

[289] G. Ilharco, M. Wortsman, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "OpenCLIP," 2021. 14

[290] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv:1907.11692*, 2019. 14

[291] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, S. Xing, G. Chen, J. Pan, J. Yu, Y. Wang, L. Wang, and Y. Qiao, "Internvideo: General video foundation models via generative and discriminative learning," *arXiv*, 2022. 14, 15, 16

[292] J. Hernandez, R. Villegas, and V. Ordonez, "Visual representation learning from unlabeled video using contrastive masked autoencoders," *arXiv*, 2023. 14

[293] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE TPAMI*, vol. 42, no. 2, pp. 502–508, 2019. 14

[294] Y. Lin, C. Wei, H. Wang, A. Yuille, and C. Xie, "Smaug: Sparse masked autoencoder for efficient video-language pre-training," in *Proceedings of ICCV*, pp. 2459–2469, 2023. 13, 14, 15, 16, 17

[295] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *Proceedings of ICML*, pp. 5178–5193, 2023. 14

[296] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S. Huang, F. Huang, and J. Zhou, "mplug-2: A modularized multi-modal foundation model across text, image and video," in *Proceedings of ICML*, 2023. 13, 14, 15, 16, 17, 18

[297] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proceedings of NeurIPS*, 2011. 14

[298] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021. 14

[299] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, J. Xu, Z. Wang, *et al.*, "Internvideo2: Scaling video foundation models for multimodal video understanding," *arXiv:2403.15377*, 2024. 14, 15, 19

[300] L. Zhao, N. B. Gundavarapu, L. Yuan, H. Zhou, S. Yan, J. J. Sun, L. Friedman, R. Qian, T. Weyand, Y. Zhao, *et al.*, "Videoprism: A foundational visual encoder for video understanding," *arXiv:2402.13217*, 2024. 14, 15, 16, 17

[301] J. Wu, Y. Jiang, Q. Liu, Z. Yuan, X. Bai, and S. Bai, "General object foundation model for images and videos at scale," in *Proceedings of CVPR*, 2024. 14

[302] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE TPAMI*, vol. 37, no. 3, pp. 630–643, 2016. 14

[303] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva-02: A visual representation for neon genesis," *arXiv:2303.11331*, 2023. 14

[304] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv*, 2023. 13, 17

[305] P. Jin, R. Takanobu, C. Zhang, X. Cao, and L. Yuan, "Chat-univi: Unified visual representation empowers large language models with image and video understanding," in *Proceedings of CVPR*, 2023. 13

[306] G. Sun, W. Yu, C. Tang, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Fine-grained audio-visual joint representations for multimodal large language models," *arXiv:2310.05863*, 2023. 13

[307] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *arXiv*, 2023. 13

[308] S. Munasinghe, R. Thushara, M. Maaz, H. A. Rasheed, S. Khan, M. Shah, and F. Khan, "Pg-video-llava: Pixel grounding large video-language models," *arXiv preprint arXiv:2311.13435*, 2023. 13

[309] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv:2205.15868*, 2022. 13, 18

[310] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, "Nüwa: Visual synthesis pre-training for neural visual world creation," in *Proceedings of ECCV*, pp. 720–736, 2022. 13

[311] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, "Video generation models as world simulators." https://openai.com/sora, 2024. 13, 14, 19

[312] Z. Yuan, R. Chen, Z. Li, H. Jia, L. He, C. Wang, and L. Sun, "Mora: Enabling generalist video generation via a multi-agent framework," *arXiv:2403.13248*, 2024. 13

[313] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv:2209.14792*, 2022. 13, 14, 18

[314] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J.-B. Huang, M.-Y. Liu, and Y. Balaji, "Preserve your own correlation: A noise prior for video diffusion models," in *Proceedings of ICCV*, pp. 22930–22941, 2023. 13, 14, 18

[315] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv:2311.15127*, 2023. 13, 14, 18

[316] W. Chai, X. Guo, G. Wang, and Y. Lu, "Stablevideo: Text-driven consistency-aware diffusion video editing," in *Proceedings of ICCV*, pp. 23040–23050, 2023. 13, 14

[317] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv:2210.02303*, 2022. 13

[318] W. Chen, J. Wu, P. Xie, H. Wu, J. Li, X. Xia, X. Xiao, and L. Lin, "Control-a-video: Controllable text-to-video generation with diffusion models," *arXiv:2305.13840*, 2023. 13

[319] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, Y. Li, T. Michaeli, *et al.*, "Lumiere: A space-time diffusion model for video generation," *arXiv:2401.12945*, 2024. 13

[320] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *Proceedings of ICML*, pp. 5156–5165, 2020. 13

[321] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023. 14

[322] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, "Fatezero: Fusing attentions for zero-shot text-based video editing," in *Proceedings of ICCV*, pp. 15932–15942, 2023. 14

[323] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of CVPR*, pp. 22563–22575, 2023. 14

[324] T.-S. Chen, C. H. Lin, H.-Y. Tseng, T.-Y. Lin, and M.-H. Yang, "Motion-conditioned diffusion model for controllable video synthesis," *arXiv:2304.14404*, 2023. 14

[325] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of CVPR*, pp. 9879–9889, 2020. 15

[326] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *Proceedings of CVPR*, pp. 23034–23044, 2023. 15, 16

[327] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *Proceedings of ECCV*, p. 1–18, 2022. 15, 16

[328] S. Ahmad, S. Chanda, and Y. S. Rawat, "Ez-clip: Efficient zeroshot video action recognition," *arXiv:2312.08010*, 2023. 16

[329] X. Huang, H. Zhou, K. Yao, and K. Han, "Froster: Frozen clip is a strong teacher for open-vocabulary action recognition," in *Proceedings of ICLR*, 2024. 16

[330] H. Akbari, D. Kondratyuk, Y. Cui, R. Hornung, H. Wang, and H. Adam, "Alternating gradient descent and mixture-of-experts for integrated multimodal perception," in *Proceedings of NeurIPS*, 2024. 16

[331] K. Ranasinghe and M. S. Ryoo, "Language-based action concept spaces improve video self-supervised learning," in *Proceedings of NeurIPS*, 2023. 16

[332] W. Lin, L. Karlinsky, N. Shvetsova, H. Possegger, M. Kozinski, R. Panda, R. Feris, H. Kuehne, and H. Bischof, "Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge," in *Proceedings of ICCV*, pp. 2851–2862, 2023. 16

[333] R. Qian, Y. Li, Z. Xu, M.-H. Yang, S. Belongie, and Y. Cui, "Multimodal open-vocabulary video classification via pre-trained vision and language models," *arXiv:2207.07646*, 2022. 16

[334] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *Proceedings of ECCV*, p. 105–124, 2022. 16

[335] X. Cheng, H. Lin, X. Wu, F. Yang, and D. Shen, "Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss," *arXiv:2109.04290*, 2021. 16

[336] Y. Ge, Y. Ge, X. Liu, D. Li, Y. Shan, X. Qie, and P. Luo, "Bridging video-text retrieval with multiple choice questions," in *Proceedings of CVPR*, pp. 16167–16176, 2022. 16

[337] Y. Ge, Y. Ge, X. Liu, J. Wang, J. Wu, Y. Shan, X. Qie, and P. Luo, "Miles: Visual bert pre-training with injected language semantics for video-text retrieval," in *Proceedings of ECCV*, pp. 691–708, 2022. 16

[338] J. Wang, Y. Ge, G. Cai, R. Yan, X. Lin, Y. Shan, X. Qie, and M. Z. Shou, "Object-aware video-language pre-training for retrieval," in *Proceedings of CVPR*, pp. 3313–3322, 2022. 16

[339] S.-V. Bogolin, I. Croitoru, H. Jin, Y. Liu, and S. Albanie, "Cross modal retrieval with querybank normalisation," in *Proceedings of CVPR*, pp. 5194–5205, 2022. 16

[340] C. Lin, A. Wu, J. Liang, J. Zhang, W. Ge, W.-S. Zheng, and C. Shen, "Text-adaptive multiple visual prototype matching for video-text retrieval," in *Proceedings of NeurIPS*, pp. 38655–38666, 2022. 16

[341] Y. Zhao and P. Krähenbühl, "Training a large video model on a single machine in a day," *arXiv:2309.16669*, 2023. 16

[342] Y. Wang, D. Gao, L. Yu, W. Lei, M. Feiszli, and M. Z. Shou, "Geb+: A benchmark for generic event boundary captioning, grounding and retrieval," in *Proceedings of ECCV*, pp. 709–725, 2022. 17

[343] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of ICCV*, pp. 1686–1697, 2021. 17

[344] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "Swinbert: End-to-end transformers with sparse attention for video captioning," in *Proceedings of CVPR*, pp. 17949–17958, 2022. 17

[345] X. Gu, G. Chen, Y. Wang, L. Zhang, T. Luo, and L. Wen, "Text with knowledge graph augmented transformer for video captioning," in *Proceedings of CVPR*, pp. 18941–18951, 2023. 17

[346] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, "Star: A benchmark for situated reasoning in real-world videos," in *Proceedings of NeurIPS*, 2021. 17

[347] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," in *Proceedings of ICLR*, 2019. 17, 19

[348] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of CVPR*, pp. 22563–22575, 2023. 18

[349] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *arXiv:2211.11018*, 2022. 18

[350] W. Wang, H. Yang, Z. Tuo, H. He, J. Zhu, J. Fu, and J. Liu, "Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation," *arXiv:2305.10874*, 2023. 18

[351] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Proceedings of NeurIPS*, pp. 22199–22213, 2022. 19

[352] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang, *et al.*, "Towards open vocabulary learning: A survey," *IEEE TPAMI*, no. 01, pp. 1–20, 2024. 19

[353] OpenAI, "Chatgpt [large language model]," 2023. Accessed: 2024-03-28. 19
[354] V. Perrone, H. Shen, A. Zolic, I. Shcherbatyi, A. Ahmed, T. Bansal, M. Donini, F. Winkelmolen, R. Jenatton, J. B. Faddoul, *et al.*, "Amazon sagemaker automatic model tuning: Scalable gradient-free optimization," in *Proceedings of SIGKDD*, pp. 3463–3471, 2021. 19
[355] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," in *Proceedings of NeurIPS*, 2023. 19
[356] I. Balažević, Y. Shi, P. Papalampidi, R. Chaabouni, S. Koppula, and O. J. Hénaff, "Memory consolidation enables long-context video understanding," *arXiv:2402.05861*, 2024. 19
[357] H. Liu, W. Yan, M. Zaharia, and P. Abbeel, "World model on million-length video and language with ringattention," *arXiv:2402.08268*, 2024. 19
[358] M. C. Schiappa, N. Biyani, P. Kamtam, S. Vyas, H. Palangi, V. Vineet, and Y. S. Rawat, "A large-scale robustness analysis of video action recognition models," in *Proceedings of CVPR*, pp. 14698–14708, 2023. 19
[359] R. Modi, V. Vineet, and Y. Rawat, "On occlusions in video action detection: Benchmark datasets and training recipes," *Proceedings of NeurIPS*, 2024. 19
[360] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proceedings of WACV*, pp. 5935–5943, 2023. 19
[361] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of ECCV*, pp. 405–421, 2020. 19
[362] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Proceedings of ECCV*, pp. 194–210, 2020. 19
[363] A. Kumar, A. Kumar, V. Vineet, and Y. S. Rawat, "Benchmarking self-supervised video representation learning," *arXiv:2306.06010*, 2023. 20
[364] G. Zara, V. G. T. da Costa, S. Roy, P. Rota, and E. Ricci, "Simplifying open-set video domain adaptation with contrastive learning," *CVIU*, p. 103953, 2024. 20
[365] S. Kareer, V. Vijaykumar, H. Maheshwari, P. Chattopadhyay, J. Hoffman, and V. Prabhu, "We're not using videos effectively: An updated domain adaptive video segmentation baseline," *arXiv:2402.00868*, 2024. 20
[366] J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," *arXiv:2312.07533*, 2023. 20

Neelu Madan is a PostDoc at the Visual Analysis and Perception lab, Aalborg University, Denmark. She completed her Ph.D. at Aalborg University in 2023 and received the 2024 Sparnord Fondens Forskningspris for her outstanding doctoral thesis. Her research interests include artificial intelligence, computer vision, machine learning, and deep learning.

Andreas Møgelmose is associate professor at the Visual Analysis and Perception lab, Aalborg University, Denmark. He received his PhD from Aalborg University in 2015 and has worked in both research and applied AI since then. His research spans a broad range of topics in the realm of computer vision, from industrial 3D scanning over human color understanding to medical image analysis.

Rajat Modi is a Ph.d. student in Computer Science at the Center For Research in Computer Vision, University of Central Florida, USA. His areas of research include Object centric representations, Neural Implicit representations, Part-Whole Hierarchies, Video Understanding and Mortal Computation.

Yogesh Rawat is an assistant professor at the Center for Research in Computer Vision, University of Central Florida, USA. He received his PhD in computer science at School of Computing, National University of Singapore. His research is focused on computer vision with special interest in video understanding, robust machine learning and multi-modal learning.

Thomas B. Moeslund leads the Visual Analysis and Perception lab at Aalborg University, the Media Technology section at Aalborg University and the AI for the People Center at Aalborg University. His research covers all aspects of software systems for automatic analysis of visual data, especially including people.

Abbreviations for Discriminative Pretraining Objective.
AGC: Attention-Guided Contrastive
BL: Box Loss
CL: Confidence Loss
Con: Cross-modal Contrastive
CTL: Contrastive Tracking Loss
ITC: Image-Text Contrastive
ITM: Image-Text Matching
CITC: Concatenated Image-Text Contrastive
CITM: Concatenated Image-Text Matching
CT: Control Task
CME: Cross-modal Moment Exploration
DVDM: Discriminative Video Dynamics Modeling
FTM: Frame-Transcript Matching
MCM: Multi-Choice Modeling
MGA: Multi-Grained Aligning
MGC: Multi-Grained Contrastive
MGL: Multi-Grained Localization
ML: Mask Loss
MMC: Multi-modal Contrastive
MTC: Multimodal Temporal Contrastive
MTRE: Multi-modal temporal relation exploration
OM-VCC: Omni-Modality Video-Caption Contrastive
OM-VCM: Omni-Modality Video-Caption Matching
VAC: Video-Audio Contrastive
PEM: Prompting Entity Modeling
SL: Semantic Loss
TMA: Tri-Modal Alignment
TOCC: Time-Order Consistency Check
TR: Temporal Reordering
VLAA: Visual Language Audio Alignment
VTM: Video-Audio Matching
VTA: Video-Text Alignment
VCC: Video Clip Contrastive
VFC: Verb-Focused Contrastive
VTC: Video-Text Contrastive
VTJ: Video-Text Joint
VTM: Video-Text Matching

Abbreviations for Generative Pretraining Objective.
AVCont: Audio Video Continuation
Captioning: Captioning Loss
CGM: Concatenated Generation Modeling
CI2VG: Compositional Image-to-video generation
CMFM: Conditioned Masked Frame Model
CMLM: Concatenated Mask Language Modeling [1]
CMLM: Conditioned Masked Language Model[2]
CS2VG: Compositional sketch-to-video generation
CVI: Compositional video inpainting
Distill: Distillation loss
FP: Frame Prediction
GVM: Generative Video Modeling
ITG: Image-text Generation
LM: Language Modeling
OM-VCG: Omni-Modality Video Caption Generation
MDM: Mask Data Modeling
MFM: Mask Frame Modeling
MFP: Mask Feature Prediction
MIM: Mask Image Modeling
MLM: Mask Language Modeling
MMM: Mask Modality Modeling
MSM: Mask Signal Modeling
MTP: Mask Then Predict
MVM: Mask Video Modeling
N(I/M/T)G: Next (Image/Motion/Text) Generation
PrefixLM: Prefix Language Modeling
TVC: Text-guided Video Completion
VI/O: Video Inpainting and Outpainting

Other Abbreviations in Paper.
DE Dual-Encoder
ED: Encoder-Decoder
FiB: Fill-in-Blank
JE: Joint-Encoder
LLMs: Large Language Models
LMMs: Large Multimodal Models
MC: Multiple Choice
ME: Mix-Encoder
MLE: Multi-Encoder
SAM: Segment Anything
STAL: Spatio-Temporal Action Localization
TAL: Temporal Action Localization
T2I: Text-to-Image
T2V: Text-to-Video
UFMs: Universal Foundation Models
V2T: Video-to-Text
VAD: Video Anomaly Detection
VideoQA: Video Question Answering
VOD: Video Object Detection
VOS: Video Object Segmentation
ViFMs: Video Foundation Models

1. This is the expansion of CMLM in method called COSA [216]
2. This is the expansion of CMLM in method called UniVL [120]