Segment Anything for Videos: A Systematic Survey

Chunhui Zhang, *Student Member, IEEE*, Yawen Cui, Weilin Lin, Guanjie Huang, Yan Rong,
Li Liu, *Member, IEEE*, Shiguang Shan, *Fellow, IEEE*

Abstract—The recent wave of foundation models has witnessed tremendous success in computer vision (CV) and beyond, with the segment anything model (SAM) having sparked a passion for exploring task-agnostic visual foundation models. Empowered by its remarkable zero-shot generalization, SAM is currently challenging numerous traditional paradigms in CV, delivering extraordinary performance not only in various image segmentation and multi-modal segmentation (*e.g.*, text-to-mask) tasks, but also in the video domain. Additionally, the latest released SAM 2 is once again sparking research enthusiasm in the realm of promptable visual segmentation for both images and videos. However, existing surveys mainly focus on SAM in various image processing tasks, a comprehensive and in-depth review in the video domain is notably absent. To address this gap, this work conducts a systematic review on SAM for videos in the era of foundation models. As the first to review the progress of SAM for videos, this work focuses on its applications to various tasks by discussing its recent advances, and innovation opportunities of developing foundation models on broad applications. We begin with a brief introduction to the background of SAM and video-related research domains. Subsequently, we present a systematic taxonomy that categorizes existing methods into three key areas: video understanding, video generation, and video editing, analyzing and summarizing their advantages and limitations. Furthermore, comparative results of SAM-based and current state-of-the-art methods on representative benchmarks, as well as insightful analysis are offered. Finally, we discuss the challenges faced by current research and envision several future research directions in the field of SAM for video and beyond.

Index Terms—Survey, Segment anything model, Video understanding, Video generation, Video editing.

I. INTRODUCTION

FOUNDATION models [1]–[3] have become a significant area of research in recent years, revolutionizing various fields such as natural language processing (NLP), computer vision (CV), and machine learning. These models are typically pre-trained on massive datasets, enabling them to learn general representations of the input data and extract meaningful features that can be further fine-tuned for specific applications.

C. Zhang is with Shanghai Jiao Tong University, The Hong Kong University of Science and Technology (Guangzhou), and also with the CloudWalk Technology Co., Ltd. E-mail: chunhui.zhang@sjtu.edu.cn.

Y. Cui is with the University of Oulu, Finland, and also with The Hong Kong University of Science and Technology (Guangzhou), China. E-mail: yawencui@oulu.fi.

W. Lin, G. Huang, Y. Rong, and L. Liu are with The Hong Kong University of Science and Technology (Guangzhou). E-mails: {wlin760, ghuang565, yrong854}@connect.hkust-gz.edu.cn, avrillliu@hkust-gz.edu.cn.

S. Shan is with the Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, and also with University of Chinese Academy of Sciences. E-mail: sgshan@ict.ac.cn.

C. Zhang, Y. Cui, W. Lin, G. Huang, and Y. Rong contributed equally to this work.

This work was done at The Hong Kong University of Science and Technology (Guangzhou).
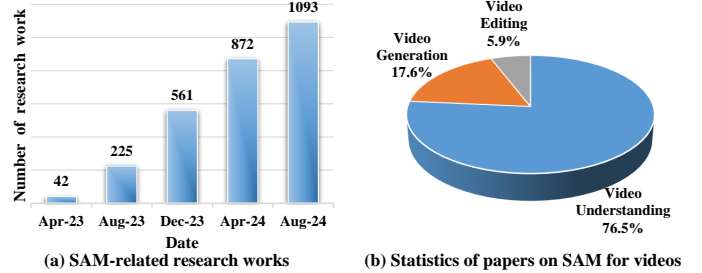
Li Liu is the corresponding author.

Fig. 1: Summarization on SAM-based works. (a) The number of SAM-related research works is rapidly increasing. (b) Video understanding dominates the research of SAM for videos.

While foundation models have primarily garnered extensive attention in NLP, their utility extends beyond that domain. In CV, researchers have been exploring the application of foundation models to enhance imaging understanding [4]–[6], object detection [7], [8], image segmentation [9], [10], and other vision-related tasks [11], [12].

One prominent example is the segment anything model (SAM) [13], which has achieved remarkable progress in exploring general and task-agnostic foundation models in the CV community. By training on over 1 billion masks on 11 million images, SAM can deliver high-quality segmentation masks based on multiple prompts (*e.g.*, points, box, and text). More importantly, SAM exhibits powerful zero-shot generalization in various segmentation tasks (*e.g.*, interactive segmentation, semantic segmentation, and panoptic segmentation), without the retraining or finetuning previously required [14]. Therefore, the emergence of SAM has led many researchers to believe that this is *"the GPT-3 moment for CV, as SAM has learned the general concept of what an object is, even for unknown objects, unfamiliar scenes (*e.g.*, underwater and cell microscopy and ambiguous cases)"* [15]. A large number of researchers have extended SAM to different fields [16]–[20]. As shown in Fig. 1(a), the number of SAM-related research works has increased significantly since April 2023[1]. The segment anything model 2 (SAM 2) [21] enhances its predecessor, SAM, by integrating a transformer framework with streaming memory, facilitating superior real-time video segmentation capabilities. Trained on the extensive and diverse segment anything video (SA-V) dataset, SAM 2 demonstrates heightened accuracy and efficiency over SAM, particularly in video tasks, and offers a robust solution for promptable visual segmentation across varied spatio-temporal contexts.

Incorporating SAM into Video Tasks. Video is an incredibly important medium in today's digital age [22]. Compared to

---

[1]https://github.com/liliu-avril/Awesome-Segment-Anything

static image and pure text, video offers strong visual representation, enhanced perception and memory, powerful storytelling capabilities, and rich interactivity, making it a more effective medium for communication and entertainment [2], [22]. The exploration of SAM in video tasks is quickly becoming a booming area of research [3], [23]–[25]. Although SAM has shown great potential in various image tasks, it still faces numerous challenges in video tasks, such as ensuring SAM's ability to consistently and coherently generate masks across lengthy video frames [16], [19] and enhancing its scalability and efficiency for handling large-scale video data [26], [27]. Most current works on video-related tasks usually employ SAM directly to achieve remarkable results of innovative applications. For a comprehensive understanding of this cutting-edge research field, as illustrated in Fig. 1(b), we conducted this survey and categorized existing works of innovative applications with SAM into three major categories (*i.e.*, video understanding, video generation, and video editing).

Unique Challenges in Videos. Compared with other tasks, *e.g.*, image and text processing, video tasks present the following unique challenges [3], [22]–[25]. **1)** Temporal information processing: video data encompasses not only spatial information but also temporal dynamics. Thus, handling video data requires considering the temporal relationships and dynamic changes. **2)** High-dimensional data: each frame of a video consists of high-dimensional data with a large number of pixels, leading to a massive amount of data that demands more computational resources and storage space. **3)** Continuity and stability: videos are generally continuous, and processing them involves considering the coherence and stability between frames to achieve reliable results in analysis and applications. **4)** Time cost: due to the substantial volume of video data, the time cost for processing video tasks is usually higher, posing greater demands on computational resources and algorithm efficiency. **5)** Action and event recognition: compared to static images, video tasks often involve recognizing actions and events, requiring models to understand and learn dynamic changes in temporal sequences. The above challenges foreshadow the extreme complexity of video tasks and enormous research opportunities [16], [22], [27].

Comparisons with Previous Surveys. Although three surveys [3], [9], [31] have been proposed for SAM, the differences between our survey and existing ones are mainly in three aspects. **1)** Previous SAM-based surveys only focus on medical image segmentation tasks [9] or roughly cover video tasks [3], [31], however, SAM for videos is a challenging and promising research topic with many innovation opportunities and potential applications [22]. This inspires us to conduct a systematic survey dedicated to this specific field (*i.e.*, SAM for videos) to benefit relevant researchers and practitioners. **2)** This survey provides an understandable and highly structured taxonomy of SAM for videos, dividing existing methods into three major categories (*i.e.*, video understanding, video generation, and video editing), which is significantly different from previous ones. **3)** A comprehensive performance evaluation, together with many new insights on SAM for videos are offered to help readers track recent advances. Additionally,

the proposed research directions are deliberate and can pave new avenues for developing foundation models in the video domain and beyond. For a comprehensive understanding of foundation models, we also refer readers to other excellent surveys for language [32]–[34], vision [22], [35], and multi-modality [1], [2].

The main contributions of this survey are threefold:

- We thoroughly review the development of SAM for videos in the foundation models era and provide a systematic survey of the latest progress in this field, which can be grouped into three major categories: video understanding, video generation, and video editing. To the best of our knowledge, this is the first systematic survey that focuses on this specific domain.
- We comprehensively compare SAM-based methods with current state-of-the-art (SOTA) methods on representative datasets for various video tasks. Importantly, our in-depth analysis about the pros and cons of these leading-edge methods can help readers choose appropriate baselines for their specific applications while delivering valuable insights on improving existing methods.
- Based on the systematic literature review and comprehensive performance evaluation, we highlight some potential future developmental trends.

The remainder of this survey is organized as follows. Section II summarizes the background knowledge, including the workflows of SAM and SAM 2, research routes, and relevant research domains. In Section III, we primarily present an overview of methods in the field of video understanding with SAM. In Section IV, we delve into the principal studies concerning video generation with SAM. In Section V, we elucidate the methods for video editing with SAM. Section VI introduces the benchmark datasets and evaluation. In Section VII, we conclude this article and highlight the potential avenues for future research.

II. PRELIMINARIES

In this section, we first briefly introduce SAM, then review three video-related research domains, including video understanding, video generation, and video editing.

A. Segment Anything Models

SAM is the segment foundation model proposed by Meta [13], as illustrated in Fig. 2(a). The pathway of SAM consists of three steps, namely task, model, and data. Inspired by large language models, tasks in SAM are usually introduced using prompt engineering [36], where a prompt is to indicate what to segment. A unique characteristic of the promptable task is that it can return a valid segmentation mask when given any segmentation prompt. The structure of SAM consists of three parts: a powerful image encoder (*i.e.*, ViT [37]); a prompt encoder, dense input, and a mask decoder (prompt-image bidirectional Transformer decoder using self-attention and cross-attention). The model is trained with focal loss [38] and dice loss [39]. Due to the insufficiency of public training data for segmentation tasks, the training-annotation iterative process is conducted in SAM by constructing a data engine to achieve

(a) SAM

(b) SAM 2

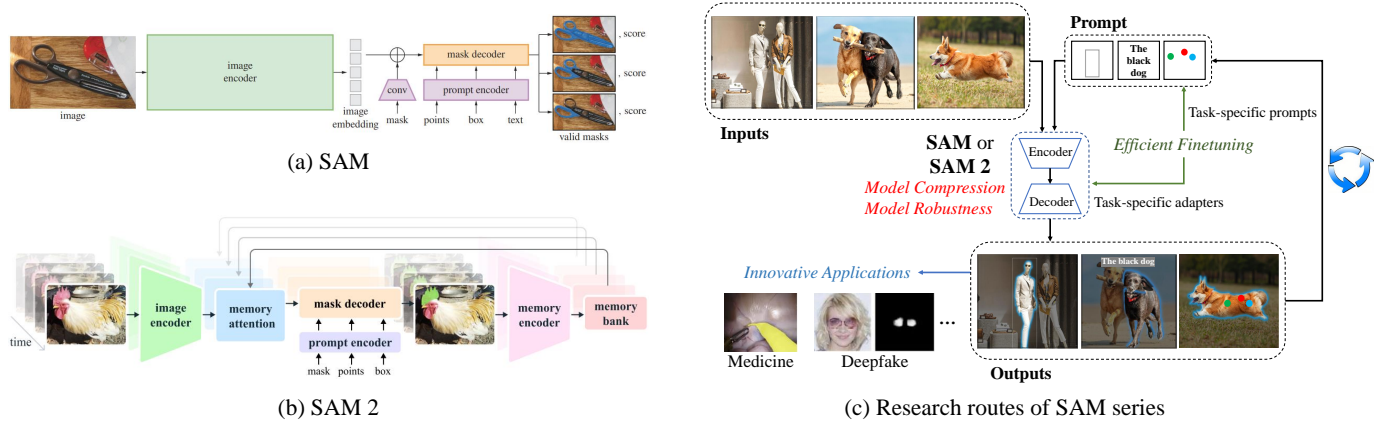(c) Research routes of SAM series

Fig. 2: Overall architectures of SAM (a) and SAM 2 (b) from the original papers [13], [21], respectively. According to the user prompts, SAM and SAM 2 can achieve interactive segmentation in images and videos. Several representative research routes for the SAM and SAM 2 models (*e.g.*, model compression [28], model robustness [29]), prompt (*e.g.*, efficient finetuning [4]), and outputs (*e.g.*, innovative applications [9], [30]) are listed in (c).

model training and dataset construction simultaneously. Benefiting from well-designed tasks, the model structure, and an extensive repository of high-quality training data, experiments demonstrate that the SAM model excels in zero-shot transfer capabilities. It has shown remarkable performance in tasks such as single-cue point segmentation, edge detection, object proposal, instance segmentation, interactive segmentation, and multi-modal segmentation (text-to-mask). Notably, the SAM model even surpasses supervised models in certain aspects.

The latest SAM 2 [21] (see Fig. 2(b)) introduces a significant evolution over its predecessor by extending its capabilities to the domain of video segmentation. SAM 2 incorporates a transformer-based architecture with a streaming memory component, enabling real-time processing of video frames. It refines the segmentation process through interactive user prompts and leverages a memory attention mechanism to retain and utilize information about the target object across frames. The SAM 2 model demonstrates improved accuracy and efficiency, requiring fewer interactions for video segmentation and outperforming SAM in both speed and accuracy for image segmentation tasks. Furthermore, the SAM 2 model is trained on the SA-V dataset, which is a substantial expansion from SAM's training data. The SA-V dataset, comprising 50.9K videos with 642.6K masklets, is not only larger but also more diverse, covering a wider range of objects and scenarios. This extensive and varied dataset has been instrumental in enhancing SAM 2's ability to segment objects in complex, real-world video content, thereby setting a new benchmark for visual segmentation tasks. The improvements in SAM 2 reflect a concerted effort to address the dynamic challenges present in video data, such as motion, deformation, and occlusion, and to provide a more generalized solution for promptable visual segmentation.

B. Research Routes of SAM

Research on SAM mainly adapts the following routes: model compression [28], ensuring model robustness [29], advancing efficient finetuning techniques [4], and developing innovative applications [9], [30] (as illustrated in Fig. 2(c)) from the perspective of methodology. In the realm of video processing, the majority of SAM research falls under the category of innovative applications, where SAM is directly applied to achieve significant outcomes. Meanwhile, a portion of the research is dedicated to refining finetuning approaches tailored for individual video tasks. Consequently, the taxonomy in this paper is based on innovative applications that SAM enables across diverse video-related challenges.

C. Related Tasks

Video Understanding. Video understanding aims to recognize and localize different actions or events appearing in the video, including (1) video recognition and (2) video localization. (1) Video recognition aims to classify the video clip or snippet into one of action or event categories. Frameworks of current works are mainly divided into two series: two-stream networks [40], [41] and single-stream RGB networks [42], [43]. This work [44] proposes a two-stream ConvNet architecture which incorporates spatial and temporal networks and demonstrates that a ConvNet trained on multi-frame dense optical flow is able to achieve very good performance in spite of limited training data. SlowFast networks [43], a one-stream framework, consists of a fast pathway operating at high frame rate and a slow pathway operating at low frame rate. (2) Video localization targets to detect and classify actions in untrimmed long videos. There are two widely used detection paradigms. The two-stage paradigm [45], [46] first localizes class-agnostic action proposal, then classifies and refines each proposal. Another one-stage paradigm [47], [48] combines localization and classification, which densely classifies each frame into actions or backgrounds.

Video Generation. Video generation aims to generate new videos from (1) the text (*i.e.*, text-to-video generation) or from (2) a single video. (1) Text-to-video generation. Early works [48]–[50] primarily generate videos in simple domains, such as moving digits or specific human actions. Recently, a series of works [51], [52] conduct VAE-based methods

for more realistic scenes. Inspired by text-to-image diffusion models, Video Diffusion Models (VDM) [53] are proposed with a space-time factorized U-Net with joint image and video data training. Make-A-Video [54] and MagicVideo [55] aim to generate videos by transferring progress from text-to-image generation. (2) Video generation from a single video. The methods on this task are divided into GAN-based methods [56], [57] and Patch nearest-neighbour methods [58]. Sinfusion [59] is the first work to utilize the capabilities of diffusion models to learn the appearance and dynamics of a single video for generating new videos.

Video Editing. Video editing usually refers to editing a video according to the textual information or an example. Video stylization is a specific type of editing task where the style provided by an example frame is propagated to the video. Existing methods can be roughly divided into (1) propagation-based methods and (2) video layering-based methods. (1) Propagation-based methods use keyframes [60], [61] to propagate edits throughout the video. This work [60] proposes a new type of guidance for SOTA patch-based synthesis, which can be applied to any type of video content. (2) Layer-based methods [62], [63] usually decompose the video into layers that are then edited. Layered neural atlases [62] map the foreground and background of a video to a canonical space, which is then operated for video editing.

III. VIDEO UNDERSTANDING WITH SAM

In this section, we primarily introduce various video understanding tasks using SAM, as shown in Fig. 3.

A. Video Object Segmentation

Video object segmentation (VOS) is a crucial task in CV for segmenting primary objects in a video. By combining with the pre-trained segmentation model SAM, recent works present great potential in VOS. We briefly summarize them into semantic, instance, panoptic, and entity levels (see Fig. 4).

1) Video Semantic Segmentation: Zhang *et al.* [20] was the first to adopt SAM for unsupervised VOS, which performs segmentation without manual annotations. Specifically, they remove the mask prediction branch in IDOL [96] to adapt it as a novel video salient object tracking method, which is to discover the salient object and spatial-temporal trajectories. Then, they adopt SAM with the generated trajectories as the prompt to obtain mask results frame by frame.

Besides, one-shot object segmentation customizing the great segmentation ability of SAM also works well in image segmentation and video segmentation. Liu *et al.* [14] presents a training-free framework, Matcher, with one-shot object segmentation. They integrate an all-purpose feature extraction model (*e.g.*, DINOv2 [10], CLIP [97], and MAE [98]) and a class-agnostic segmentation model (*i.e.*, SAM) with three operations to realize the controllable masks generation. Following it, Zhang *et al.* [19] introduce a training-free personalization SAM, named PerSAM, to segment only the user-provided object with SAM. Specifically, they first obtain a location confidence map for the target object with the user-provided image and mask. Then, based on the confidence scores, they proposed

target-guided attention and target-semantic prompting to aid SAM's decoder for personalized segmentation. Additionally, they provide a finetuning variant PerSAM-F with only 2 parameters within 10 seconds to alleviate the mask ambiguity issue. Both of the methods can be used in image and video object segmentation in the frame-by-frame setting.

Except for them, Chang *et al.* [64] adopting SAM as a post-processing technique for semantic segmentation in the PVUW2023 VSS track. Zhou *et al.* [65] propose a novel moving object segmentation (MOS) dataset, named DSEC-MOS, with high temporal resolution and low-latency information on the changes of scenes to promote the research on MOS.

2) Video Instance Segmentation: To solve the problems of coarse mask boundaries and incorrect predictions on SAM, Ke *et al.* [16] propose HQ-SAM, which equips SAM with the ability to segment any object more accurately. To be specific, they introduce a lightweight High-quality Output Token to replace the original SAM's output token and a Global-local Feature Fusion to fuse the global semantic context and the local boundary details. They fix the pre-trained model parameters to keep the original performance of SAM and only train a few parameters of the introduced components on their composed dataset with 44K fine-grained masks.

3) Video Panoptic Segmentation: One essential challenge for the end-to-end video segmentation model is poor performance in large-vocabulary settings. In the large-vocabulary dataset VIPSeg [99], a recent work [100] achieves only 26.1 in terms of video panoptic quality score. Cheng *et al.* [66] state that the increasing number of classes and scenarios makes it difficult to conduct end-to-end training with good performance. Therefore, they propose a decoupled video segmentation approach (DEVA) with task-specific image-level segmentation and class/task-agnostic bi-directional temporal propagation. Specifically, SAM is used for image-level segmentation with universal data training containing outside-the-target-domain data. With the first segmented frame, they denoise the error with a few frames in the near future to reach a consensus as the output segmentation. Then, XMem [101] is adapted as the temporal propagation model to propagate the segmentation to subsequent frames. The extensive experiments on VIPSeg validate its effectiveness on large-scale video panoptic segmentation.

4) Video Entity Segmentation: The in-the-wild setting of the image/video segmentation task is a big challenge for the existing methods, where no restriction is set on domains, classes, image resolution, and quality [67]. Although entity segmentation is designed to segment unseen categories in the training set, the lack of entity segmentation datasets makes it difficult to develop well on this task. To fill this gap, Qi *et al.* [67] construct a high-quality large-scale entity segmentation dataset, named EntitySeg. The dataset contains 33,227 images with high-quality annotated masks on multiple domains and diverse resolutions, allowing the evaluation of the models' generalization and robustness. They benchmark the existing models and find that they cannot well accommodate the proposed dataset. Therefore, they further propose the CropFormer [67] framework to solve the problem.
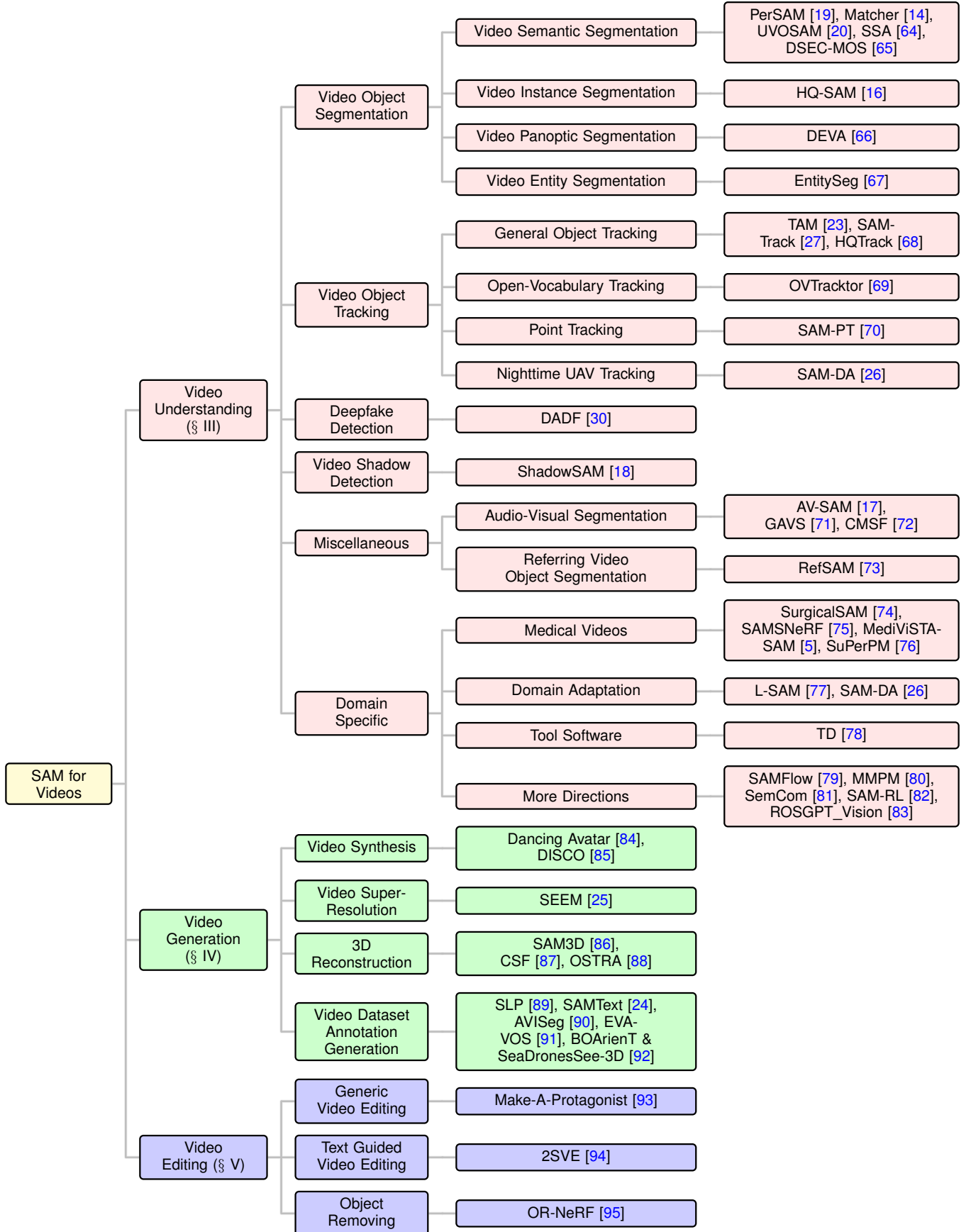
Fig. 3: Taxonomy of research works on SAM for videos. Due to space considerations, we merely list some representative methods for each video-related task here.

B. Video Object Tracking

Video object tracking (VOT) is a fundamental task in CV. We divide VOT methods using SAM into four groups: (1) general object tracking, (2) open-vocabulary tracking, (3) point tracking, and (4) nighttime unmanned aerial vehicle (UAV) tracking. (1) Recently, the strong segmentation ability of SAM enhances the perception of objects and allows better development in general object tracking. Yang *et al.* [23] propose the training-free track anything model (TAM) based on SAM to achieve high-performance interactive tracking and segmentation in videos. Specifically, they first use SAM to get an initial mask of objects, where the user can choose the target object by a click or modify the mask. Then, they adopt XMem to perform VOS on the following frames with the user-selected mask. To avoid the issue that XMem segments more coarsely over time, they use SAM again to refine it. Cheng *et al.* [27] propose SAM-Track to segment and track any object in a video. They incorporate SAM to obtain segments, Grounding-DINO to understand natural language, and DeAOT [102] for tracking. In the VOTS2023 challenge, Zhu *et al.* [68] won 2nd place to achieve high-quality VOT with their proposed HQTrack. Specifically, the framework implements the improved variants of DeAOT and SAM (*i.e.*, HQ-SAM [16]) for multi-object segmentation and mask refining, respectively. A similar idea of combining SAM and DeAOT is seen in the 1st place solution of TREK-150 object tracking challenge [103]. They introduce MSDeAOT as an improved variant of DeAOT by replacing the bounding box with masks in the reference frame and feeding the mask and frames into the VOS model. (2) Chu *et al.* [69] utilize SAM as the segmenter along with an open-vocabulary object detector and an optical flow estimation to build a zero-shot open-vocabulary visual tracking framework OVTracktor. (3) SAM-PT [70] was proposed to utilize the sparse point propagation of VOS. Taking a video with point annotations in the first frame as input, SAM-PT can achieve strong zero-shot performance with a point tracker to generate the trajectories as prompts and SAM to output predicted masks. The predicted masks are also used to reinitialize and get rid of the unreliable points. (4) Yao *et al.* [26] utilize SAM for the field of real-time nighttime UAV tracking to accurately locate the potential object and determine high-quality target domain training samples from the night-time images.

C. Deepfake Detection

In a recent investigation, Lai *et al.* [30] delved into evaluating the performance of SAM and its variants in the context of deepfake detection and localization, marking the first attempt to assess these methods for this specific task. The researchers noted that existing approaches, which utilize LoRA [104], SAM adapter [4], and learnable prompt [105] to fine-tune SAM on downstream tasks, often yielded unsatisfactory results, particularly in terms of face forgery localization. This inadequacy was attributed to their limited capacity in modeling both local and global contexts for forgery.

To tackle these challenges, Lai *et al.* [30] proposed an innovative framework, named detect any deepfakes (DADF),

building upon SAM. Specifically, they introduced a multi-scale adapter within SAM designed to capture short- and long-range forgery contexts, facilitating efficient finetuning. Additionally, a reconstruction guided attention module was introduced to enhance forged traces and boost the model's sensitivity toward forgery regions. The proposed method exhibited SOTA performance in both forgery detection and localization.

D. Video Shadow Detection

The detection of video shadows plays a crucial role in various applications, including object detection [106], image segmentation [107], and virtual reality scene generation [108]. However, the challenge lies in the limited availability of training data, posing difficulties for the generalization capability of existing deep neural network based methods. These limitations can lead to prediction errors accumulating during video propagation [109].

In particular, when applying SAM to single-frame shadow detection, SAM tends to categorize shadows as part of the background [18]. This introduces a nontrivial challenge in using SAM for shadow detection, as it requires bridging the gap between natural objects and complex shadows. To address this challenge, Wang *et al.* [18] introduced ShadowSAM, a straightforward yet effective framework designed for finetuning SAM specifically for shadow detection. Additionally, by adopting a long and short-term attention mechanism, they extended its capabilities for efficient video shadow detection.

E. Miscellaneous

1) Audio-Visual Segmentation: Recently, SAM was applied in audio-visual localization and segmentation [17], [71]. Both studies focus on overcoming challenges associated with audio-visual localization and segmentation, particularly addressing the inherent misalignment between audio and various objects in the video.

In [17], the authors address this challenge by introducing AV-SAM, a method that learns audio-aligned visual features for each mask prompt from the video. This facilitates the guidance of mask generation in SAM through pixel-wise audio-visual fusion. The approach utilizes audio features and visual features from the pre-trained image encoder in SAM to aggregate cross-modal representations. Conversely, Wang *et al.* [71] present an encoder-prompt-decoder paradigm to tackle issues related to data scarcity and varying data distribution. Leveraging abundant knowledge from pre-trained models, they introduce a semantic-aware audio prompt to assist the visual foundation model in focusing on sounding objects. Simultaneously, this approach encourages the reduction of the semantic gap between visual and audio modalities. Furthermore, Bhosale *et al.* [72] propose CMSF, a method leveraging audio cues to generate audio tags and subsequently proposing segmentation masks. These recent advancements underscore the versatility of SAM in addressing intricate tasks related to audio-visual processing.

2) Referring Video Object Segmentation: Despite SAM gaining widespread attention for its impressive performance in image segmentation, a study discussed in [73] highlights

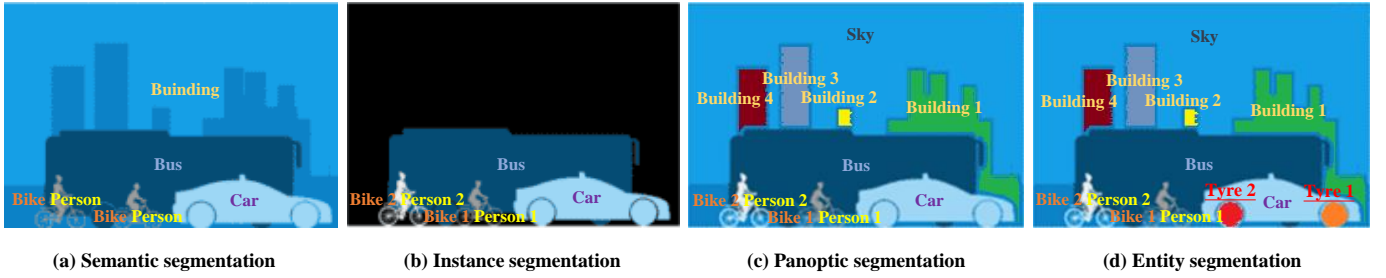| (a) Semantic segmentation | (b) Instance segmentation | (c) Panoptic segmentation | (d) Entity segmentation |

Fig. 4: Concepts comparison of four prevalent visual segmentation tasks, including semantic, instance, panoptic, and entity segmentation. (a) For semantic segmentation, the same textures or categories are assigned the same class labels. (b) Instance segmentation only focuses on the foreground, and different objects in the same category are assigned different instance identities. (c) In panoptic segmentation, each pixel is assigned a semantic label and a unique instance identifier. (d) Entity segmentation [67] requires segmenting unseen categories in the training set, *e.g.*, "tyre".

SAM's limitations in the realm of referring video object segmentation (RVOS). This limitation stems from the need for precise user interactive prompts and a constrained understanding of different modalities, such as language and vision.

In a concerted effort to effectively tailor SAM for RVOS in an end-to-end manner and fully unleash its potential for video segmentation and multi-modal fusion, Li *et al.* conducted a groundbreaking study [73]. They delved into SAM's potential for RVOS by integrating multi-view information from diverse modalities and successive frames at different timestamps. The authors introduced RefSAM, a novel approach that utilizes lightweight modules and an efficient finetuning strategy to align and fuse language and vision features in an end-to-end learning fashion. Additionally, they designed a hierarchical dense attention module to exploit diverse levels of visual and textual features, thereby facilitating effective cross-modal segmentation of objects with varying sizes.

F. Domain Specific

1) Medical Videos: SAM also contributes to the analysis of medical videos. Regarding the two problems with naive pipeline of SAM (*i.e.*, the domain gap and the dependency on precise point or box locations), SurgicalSAM [74] introduces a novel end-to-end efficient finetuning approach for SAM, and the objective is to seamlessly incorporate surgical-specific information with SAM's pre-trained knowledge for enhancing overall generalization capabilities. This work [110] comprehensively explores different scenarios of robotic surgery and evaluates SAM's robustness and zero-shot generalizability. SAMSNeRF [75] combines SAM and neural radiance field (NeRF) techniques, which generates accurate segmentation masks of surgical tools using SAM and then guides the refinement of the dynamic surgical scene reconstruction by NeRF. Fillioux *et al.* [111] evaluate SAM's performance on processing patient-derived organoids microscopy frames. MediViSTA-SAM [5] is the first study on adapting SAM to video segmentation. SuPerPM [76] is a large deformation-robust surgical perception framework, which utilizes SAM to segment tissue regions from the background.

2) Domain Adaptation: Recently, researchers utilized SAM to enhance the generalization ability of the model on target domain, especially in situations where the quality and quantity of data in the target domain are less than ideal. Bonani *et al.* [77] utilized SAM to provide a regularization signal for real data and introduced an invariance-variance loss structure. This structure is defined for self-supervised learning on unlabeled target domain data, facilitating the robustness of domain adaptation ability for semantic segmentation networks. Yao *et al.* [26] proposed SAM-DA, a SAM-powered domain adaptation framework designed for real-time nighttime UAV tracking. They introduced an innovative SAM-driven method to expand target domain training samples, which generates a substantial quantity of high-quality training samples for the target domain from each nighttime image, enabling one-to-many sample generation. This approach significantly augments both the quantity and quality of target domain training samples, thereby providing improved data support for domain adaptation.

3) Tool Software: Hsieh *et al.* [78] explored the possibility of leveraging tool documentation, as opposed to demonstrations, for instructing large language models (LLMs) on the utilization of new tools. The article [78] demonstrated that the use of tool documentation empowered LLMs to employ SAM in a zero-shot manner, eliminating the need for training or finetuning. Of equal significance, the article showcased the potential of employing tool documentation to enable novel applications. One such illustration involved the amalgamation of GroundingDino [7] and SAM, resulting in the creation of Grounded-SAM [112]—a model proficient in generating text grounded in visual content, showcasing its capabilities to derive meaningful textual information from images.

4) More Directions: Several studies have applied SAM in various applications, spanning optical flow estimation [79], robotics [80], [83], [113], reinforcement learning (RL) for video games [82], and semantic communication [81].

To address the challenge of "fragmentation" in optical flow estimation, Zhou *et al.* [79] employed SAM as an image encoder, providing optical flow estimation with richer and higher-level contextual features. This strategy mitigates the model's tendency to focus exclusively on local and low-level cues. In the work by Yang *et al.* [80], SAM was used to generate segmentation masks for objects, providing the model with rich semantic, geometric, and shape priors. This, in turn, assists robots in perceiving object poses and determining

(a) Audio-visual segmentation masks generation   (b) Text guided video editing
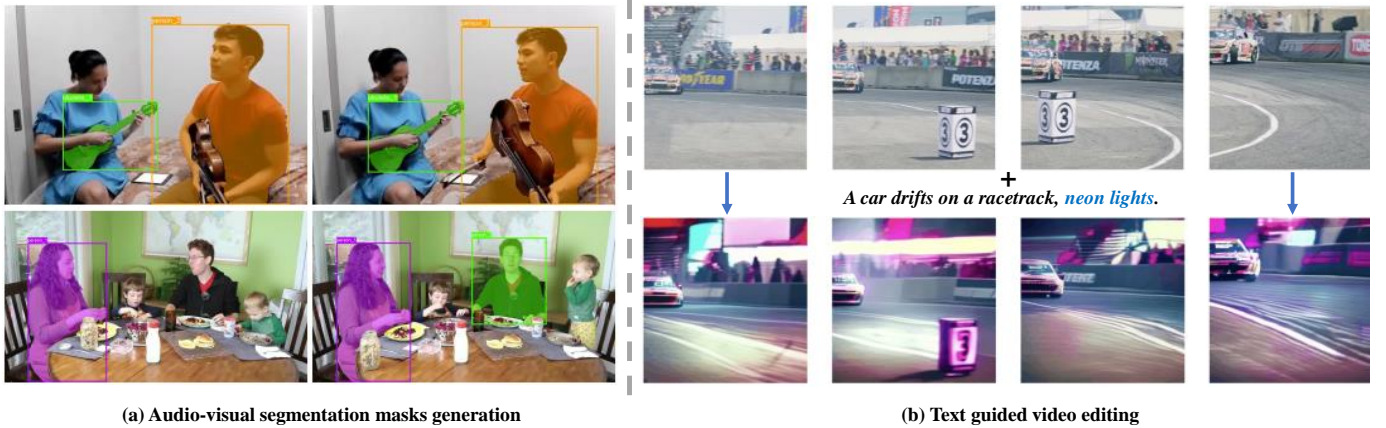
Fig. 5: Examples of video masks generation results with (a) AVISeg [90] and video editing results with (b) 2SVE [94].

grasp points. Similar ideas are also evident in [83], [113]. In [82], the authors enhanced the original pixel input using SAM, aiming to improve the performance of RL agents in Atari video games. Despite the observed improvement in the game-playing performance of the RL agent, finding a suitable balance between performance enhancement and computational cost remains an ongoing exploration. Additionally, Raha *et al.* [81] proposed a novel semantic communication framework based on SAM, efficiently transmitting sequential images or videos while preserving the original content unchanged.

IV. VIDEO GENERATION WITH SAM

In this section, we divide video generation with SAM into four groups and provide detailed reviews for each: video synthesis (*e.g.*, dance generation) [84], [85], video super-resolution (*i.e.*, generating more detailed and visually appealing videos from low-resolution versions) [25], 3D reconstruction (*e.g.*, reconstruction and segmentation of 3D objects and producing point-level semantic labels for 3D point cloud) [86], [88], and video dataset annotation generation (*e.g.*, bounding boxes and masks generation) [24], [89]–[92]. The taxonomy details of video generation with SAM is illustrated in Fig. 3.

A. Video Synthesis

SAM has been recently employed in two works focusing on dance video synthesis [84], [85]. In the Dancing Avatar project, SAM is employed to generate contextually appropriate background images for human motion videos, following textual specifications and using image inpainting techniques. This method leads to the creation of distinct pose-guided images for various poses, resulting in the generation of human area masks. In the research presented in [84], SAM ensures a consistent background throughout the human motion image sequence, effectively separating the human foreground from the background [85]. SAM's outstanding performance in these projects, along with its contributions to other modules, has played a pivotal role in achieving an impressive result.

B. Video Super-Resolution

The main challenge in video super-resolution (VSR) lies in handling large motions in input frames, making it challenging

to accurately aggregate information from multiple frames. However, according to literature [25], existing methods overlook valuable semantic information that could significantly enhance results, and flow-based approaches heavily depend on the accuracy of flow estimates, which may be imprecise for two low-resolution frames.

In [25], a robust and semantic-aware prior for improved VSR was investigated by leveraging the SAM. To incorporate the SAM-based prior, the study proposed the SAM-guidEd refinEment Module (SEEM), a simple yet effective module enhancing both alignment and fusion procedures through the utilization of semantic information. This lightweight plug-in module is designed not only to leverage attention mechanisms for generating semantic-aware features but also to be easily integrated into existing methods. Specifically, SEEM was applied to two representative methods, EDVR and BasicVSR, resulting in consistently improved performance with minimal implementation effort across three widely used VSR datasets: REDS, Vid4, and Vimeo-90K [114]. Importantly, SEEM was found to enhance existing methods efficiently, providing increased flexibility in adjusting the balance between performance and the number of training parameters.

C. 3D Reconstruction

Recent research has explored leveraging SAM's robust generalization and transfer capabilities to extend its application from 2D image segmentation to tasks related to 3D reconstruction and segmentation. These efforts bring forth a fresh perspective and methodology for understanding and reconstructing 3D scenes. Based on the approaches of converting the 2D segmentation masks into 3D masks, existing methodologies can be broadly classified into two categories: tracking-based approaches [88] and projection-based approaches [86], [87].

Tracking-based Approaches. Xu *et al.* [88] introduced OS-TRA, an open-source one stop 3D target reconstruction and multilevel segmentation framework. Within this framework, SAM is employed to segment the first frame of a video. Subsequently, the authors use VOT algorithms to generate continuous masks for video frames. In this process, SAM is also used to complementarily correct tracking errors. Finally, 3D reconstruction methods are applied to reconstruct

labeled 3D objects or multiple components. Experimental results demonstrate that OSTRA can support common 3D object models, including point clouds, meshes, and voxels. Meanwhile, in complex scenes with intricate structures and occlusions, OSTRA outperforms manual segmentation.

Projection-based Approaches. The procedural framework of projection-based approaches involves three key steps: firstly, applying SAM to segment RGB images of the 3D scene; secondly, projecting 2D segmentation masks into 3D space; and finally, generating 3D semantic pseudo labels. Yang *et al.* [86] introduced SAM3D, a 3D point cloud segmentation framework based on SAM. In the step of generating 3D semantic pseudo labels, SAM3D incorporates a bottom-up merging method, iteratively combining 3D masks from different frames, thereby consolidating the masks from two adjacent frames into a singular representation. Diverging from the aforementioned merging method [86], Dong *et al.* [87] proposed a cumulative semantic fusion (CSF) method. This approach integrates 3D segmentation results from various perspectives through a voting mechanism to generate 3D pseudo-semantic labels. Furthermore, the CSF framework addresses diverse scenarios with three segmentation strategies: a CLIP-based LSeg [115] strategy, a SAM with Grounding-DINO [7] strategy, and a SAM with 2D sparse point annotations strategy.

D. Video Dataset Annotation Generation

Due to the high cost associated with annotating videos in specific domains, many datasets lack effective labels, particularly at the pixel level. Some studies have harnessed the advantage of SAM to design systems for automatically annotating video data, providing adaptability to various scenes and objects. Balaban *et al.* [89] proposed a semantic label propagation (SLP) system that integrates SAM and structure from motion (SfM) for automatic video data annotation. SAMText [24] is a scalable mask annotation pipeline capable of rapidly generating fine mask annotations for scene text images or video frames at scale. EVA-VOS [91] employs an intelligent agent to iteratively predict frames to annotate and the annotation types to use, establishing a human-in-the-loop annotation framework for video object segmentation. Experimental results indicate that EVA-VOS's annotation method achieves segmentation quality close to human consistency and is 3.5 times faster than traditional annotation methods.

Some studies have also leveraged SAM to introduce innovative datasets [92]. SAMText-9M [24] is a novel large-scale video text detection dataset, comprising over 2,400 video clips and more than 9 million segmentation masks. In contrast to data annotated with text position using quadrilateral bounding boxes, SAMText-9M utilizes detailed segmentation masks for text annotation. In the domain of audio-visual instance segmentation, Guo *et al.* [90] constructed the first audio-visual instance segmentation dataset, AVISeg (see Fig. 5(a)), which includes 1,258 videos, 78,665 frames, 281 objects, 15,355 masks, and 26 categories. Additionally, AVISeg can be applied to various multi-modal video understanding tasks such as video editing, virtual reality, and robot navigation.

V. Video Editing with SAM

In this section, we detail the video editing algorithms using SAM that are divided into three groups: generic video editing, text guided video editing, and object removing. The taxonomy relations of video editing with SAM are illustrated in Fig. 3.

Generic Video Editing. Make-A-Protagonist [93] proposes a framework for generic video editing with both visual and textual clues. It leverages multiple pre-trained experts to process source video and target visual/textual clues. Then, all the information is put into the proposed visual-textual-based video generation model with mask-guided denoising sampling to generate the desired output. SAM plays a key role in segmenting the protagonist based on the text description and masking out the background in the reference image.

Text Guided Video Editing. Wu *et al.* [94] introduces a new dataset (TGVE) that contains 76 videos with 4 prompts each for text-guided video editing (see Fig. 5(b)). Based on the TGVE dataset, the competition workshop was held at CVPR 2023. The winning method Two-Stage Video Editing (2SVE) incorporates many pre-trained models such as SAM, OpenCLIP [116] and ControlNet [117]. The target segment process is based on the SAM and OpenCLIP models. It works as follows: SAM predicts the masks of the input frame automatically. Then, OpenCLIP converts the masks into embeddings and calculates similarity with the text embeddings to select the target mask for the next steps. The 2SVE method consists of two stages, as its name indicates. The first stage uses ControlNet to edit the foreground, background, and structure of the input video. The second stage uses a diffusion model trained on the MSVD [118] dataset to edit the style and appearance of the output video. ControlNet and diffusion model take the target masks from the target segment process as guidance in both stages.

Object Removing. Researchers also leverage the strong prompt segmentation ability of SAM to remove objects in 3D scenes. OR-NeRF [95] proposes a novel object-removing pipeline using either points or text prompts on a single view and ensuring multiview consistency and plausible completion after deletion. OR-NeRF consists of two stages: multiview segmentation and scene object removal. For the input in the first stage, the model either uses point prompts directly or converts the input text information into point prompts via Ground-SAM. The model uses SAM to predict the mask of images from all viewing angles based on prompts and uses LaMa to obtain color and depth priors. Then, using NeRF to reconstruct the scene after removal. Based on their one-step multiview segmentation method, which leverages SAM's strong power, it achieves better removal quality and requires less time than previous methods.

VI. Performance Evaluation

In this section, we introduce the benchmark datasets, evaluation metrics, and comparative results of current SOTA and SAM-based methods across different video tasks.

TABLE I: Performance evaluation of different VOS tasks, including video semantic segmentation (DAVIS 2017 val [119] and DAVIS 2016 val [120] datasets), zero-shot open-world video instance segmentation (UVO [121]), and zero-shot video instance segmentation (HQ-YTVIS [122]). The best results are marked in **bold**.

Video Semantic Segmentation										
Method	Venue	SAM-Based	Training Data	Prompt	DAVIS 2017 val			DAVIS 2016 val		
					$J\&F$	J	F	$J\&F$	J	F
XMem [101]	ECCV 2022	✗	Video [119], [123]	-	**87.7**	**84.0**	**91.4**	**92.0**	**90.7**	**93.2**
PerSAM [19]	arXiv 2023	✔	Training-free	Mask	60.3	56.6	63.9	-	-	-
PerSAM-F [19]	arXiv 2023	✔	Image [124]	Mask	71.9	69.0	74.8	-	-	-
UVOSAM [20]	arXiv 2023	✔	Video [119]	Trajectories	78.9	75.5	82.0	-	-	-
Matcher [14]	arXiv 2023	✔	Training-free	Point/Center/Box	79.5	76.5	82.6	86.1	85.2	85.2
GT Box+SAM [13]	ICCV 2023	✔	Training-free	Box	87.3	83.5	91.0	-	-	-
Zero-shot Open-world Video Instance Segmentation										
Method	Venue	SAM-Based	Training Data	Prompt	UVO					
					AP_B^{strict}	AP_{B75}^{strict}	AP_{B50}^{strict}	AP_B	AP_{B75}	AP_{B50}
DTM [125]	arXiv 2023	✗	Image [8], [126], [127]+Video [121]	-	-	-	-	27.6	**29.2**	**40.6**
SAM [13]	ICCV 2023	✔	Image [13]	Box	8.6	3.7	25.6	17.3	14.4	37.7
HQ-SAM [16]	arXiv 2023	✔	Image [16]	Box	**9.9**	**5.0**	**28.2**	18.5	16.3	38.6
Zero-shot Video Instance Segmentation										
Method	Venue	SAM-Based	Training Data	Prompt	HQ-YTVIS					
					AP^B	AP_{75}^B	AP_{50}^B	AP^M	AP_{75}^M	AP_{50}^M
VMT [122]	ECCV 2022	✗	Video [122]	-	**44.8**	**43.4**	-	**64.8**	70.1	-
SAM [13]	ICCV 2023	✔	Image [13]	Box	30.2	19.1	72.9	60.7	68.1	90.5
HQ-SAM [16]	arXiv 2023	✔	Image [16]	Box	34.0	24.3	**79.5**	63.6	**70.5**	**91.1**

A. Evaluation of Video Object Segmentation Approaches

Datasets. DAVIS 2016 [120] and DAVIS 2017 [119] are two widely used datasets for VOS. DAVIS 2016 contains 50 videos with a total of 50 annotated object instances. The dataset is split into 30 videos for training and 20 videos for validation. DAVIS 2017 is an extension of [120], with toltaling 150 videos and 376 annotated object instances. The test and validation sets of this dataset both contain 30 videos. Unidentifed Video Objects (UVO) [121] is a large-scale dataset for open-world object segmentation in videos. HQ-YTVIS [122] is a video dataset for high-quality video instance segmentation, including 1,678 videos for the training set, 280 videos for the validation set, and 280 videos for the test set.

Evaluation Metrics. In the evaluation of the video semantic segmentation task, the commonly used metrics are region ($\mathcal{J}$) and contour ($\mathcal{F}$) measures proposed in DAVIS 2016 [120], and $\mathcal{J}\&\mathcal{F}$ metric proposed in DAVIS 2017 [119] by calculating the mean of region similarity and contour accuracy over all object instances. The boundary AP_B and stricter AP_B^{strict} [16] are adopted to assess the mask quality for the UVO dataset. For the HQ-YTVIS dataset [122], the evaluation metrics include the standard tube mask AP^M and Tube-Boundary AP^B.

Results Comparison. For VOS, we mainly conduct comparison on video semantic segmentation and video instance segmentation as there are numbers of methods for benchmarking. In Tab. I, we present VOS performance of current SOTA and SAM-based methods on DAVIS 2017 val [119], DAVIS 2016 val [120], UVO [121], and HQ-YTVIS [122] datasets.

The main observations are as follows: **1)** The SAM-based methods have significant performance gaps compared to the current SOTA methods designed for specific video segmentation tasks. This is because many SAM-based methods directly utilize SAM pre-trained on image data to enhance the ability of object segmentation, ignoring the importance of end-to-end training and finetuning for inherently complex video tasks.

For instance, on DAVIS 2017 val [119], the best SAM-based video semantic segmentation method (Matcher [14]) is training free, achieving a $\mathcal{J}\&\mathcal{F}$ score of 79.5%, while Xmen [101] obtains a $\mathcal{J}\&\mathcal{F}$ score of 87.7%. Xmen is a long-term VOS method with multiple feature memory stores. This highlights the importance of memory mechanism for ensuring temporal consistency and coherence of the predicted masks across long videos. **2)** The high-quality of prompt is crucial for the zero-shot generalization ability of SAM. One example is that "GT Box+SAM" [13] uses reliable ground-truth bounding boxes as prompts and achieves the second best results on DAVIS 2017 val. **3)** The models (*e.g.*, XMem [101] and VMT [122]) trained on video data exhibit significant advantages over the models trained solely on image data, and the utilization of multi-modal data (*e.g.*, video and image) [125] often leads to improved performance.

B. Evaluation of Video Object Tracking Approaches

Datasets. VOTS2023 [128] is the first dataset to merge short-term and long-term, as well as single-target and multi-target tracking, with dense mask annotations. It contains 144 videos and 341 targets in total. TREK-150 [129] is a dataset used to evaluate visual object tracking in the first person vision. It includes 150 videos with 97K densely annotated bounding boxes. NUT-L [26] is a long-term nighttime UAV tracking benchmark consisting of 43 videos and 95,274 frames. YouTube-VOS 2018 [123] is a large VOS dataset composed of 3,252 YouTube video clips and 133,886 object annotaions. It consists of 2,796 videos for the training set, 134 videos for the validation set, and 322 videos for the test set. Recently, VOS datasets such as YouTube-VOS 2018 [123], DAVIS 2016 [120], and DAVIS 2017 [119] are also adopted to evaluate VOT algorithms.

Evaluation Metrics. For the common VOT datasets (VOTS2023 and NUT-L), five popular evaluation metrics (*i.e.*, success rate (AUC), accuracy (A), robustness (R), precision

TABLE II: Performance evaluation on six VOT datasets (VOTS2023 [128], TREK-150 [129], NUT-L [26], DAVIS 2017 test [119], DAVIS 2016 val [120], and YouTube-VOS 2018 val [123]). We summarize current SOTA and SAM-based methods. The best results are marked in **bold**.

Video Object Tracking					
Method	SAM-Based	Initialization	VOTS2023		
			AUC	A	R
DMAOT [128]	✗	Mask	**63.6**	75.1	**79.5**
HQTrack [68]	✔	Mask	61.5	**75.2**	76.6
Method	SAM-Based	Initialization	TREK-150 test set		
			MSE	OPE	HOI
LTMU-H-IJCV [130]	✗	Box	54.3	50.5	65.7
MSDeAOT [103]	✔	Box	**73.4**	**75.5**	**77.1**
Method	SAM-Based	Initialization	NUT-L		
			AUC	$nPre$	Pre
UDAT [131]	✗	Box	37.7	43.4	49.8
SAM-DA [26]	✔	Box	**43.0**	**49.2**	**56.4**
Method	SAM-Based	Initialization	DAVIS 2017 test		
			$J \& F$	J	F
SwinB-DeAOT-L [102]	✗	Mask	**82.8**	**78.9**	**86.7**
TAM [23]	✔	Click	73.1	69.8	76.4
SAM-Track [27]	✔	Click	79.2	75.3	83.1
Method	SAM-Based	Initialization	DAVIS 2016 val		
			$J \& F$	J	F
SwinB-DeAOT-L [102]	✗	Mask	**92.9**	**91.1**	**94.7**
TAM [23]	✔	Click	88.4	87.5	89.4
SAM-Track [27]	✔	Click	92.0	90.3	93.6
Method	SAM-Based	Initialization	YouTube-VOS 2018 val		
			$J \& F$	J	F
STCN [132]	✗	Mask	**83.0**	**81.9**	**86.5**
OVTracktor [69]	✔	Detected Mask	62.2	65.9	69.4
PerSAM-F [19]	✔	Mask	71.9	69.0	74.8
SAM-PT [70]	✔	Query Points	76.3	73.6	78.9
SAM-PT-reinit [70]	✔	Refining Points	76.6	74.4	78.9
HQ-SAM-PT [16]	✔	Query Points	77.2	74.7	79.8
HQ-SAM-PT-reinit [16]	✔	Refining Points	77.0	77.0	79.2

TABLE III: Performance evaluation of other video understanding tasks, including deepfake detection, video shadow detection, miscellaneous (*i.e.*, audio-visual segmentation and RVOS), and domain specific (*i.e.*, optical flow estimation and 3D point cloud segmentation) methods.

Deepfake Detection						
Method	SAM-Based	FaceForensics++				
		DF	F2F	FS	NT	Average
Locate [133]	✗	97.25	94.46	97.13	84.63	93.36
DADF [30]	✔	**99.02**	**98.92**	**98.23**	**87.61**	**95.94**
Video Shadow Detection						
Method	SAM-Based	ViSha				
		MAE	F_β	IoU	SBER	NBER
Liu *et al.* [134]	✗	0.029	0.793	0.640	16.26	1.44
ShadowSAM [18]	✔	**0.024**	**0.813**	**0.661**	**25.21**	**1.13**
Audio-Visual Segmentation						
Method	SAM-Based	AVSBench-V1S		AVSBench-V1M		
		mIoU	F-score	mIoU	F-score	
AUSS [135]	✗	**89.4**	**94.2**	63.5	75.2	
AV-SAM [17]	✔	40.8	56.6	-	-	
CMSF [72]	✔	58.0	67.0	34.0	44.0	
GAVS [71]	✔	80.1	90.2	**63.7**	**77.4**	
Referring Video Object Segmentation						
Method	SAM-Based	Ref-DAVIS17			Ref-YouTube-VOS	
		$J \& F$	J	F	$J \& F$	J	F
ReferFormer [136]	✗	61.1	58.1	64.1	**64.9**	**62.8**	**67.0**
RefSAM [73]	✔	**66.1**	**62.9**	**69.3**	55.1	53.9	56.3
Medical Videos						
Method	SAM-Based	EndoVis2017			EndoVis2018	
		cIoU	IoU	mcIoU	cIoU	IoU	mcIoU
MATIS Full [137]	✗	71.36	66.28	41.09	**84.26**	79.12	54.04
SurgicalSAM [74]	✔	69.94	**69.94**	**67.03**	80.33	**80.33**	**58.87**
SAM Box [110]	✔	**88.20**	-	-	81.09	-	-
Method	SAM-Based	CAMUS			In-house Data	
		Dice	dH	dA	Dice	dH	dA
SwinUNETR [138]	✗	94.0	5.02	1.32	87.8	13.98	5.88
MediViSTA-SAM [5]	✔	**96.0**	**4.25**	**0.74**	**91.0**	**11.03**	**3.26**
Method	SAM-Based	EndoNeRF			SuPer Data	
		PSNR	SSIM	LPIPS	V1	T1	T2
EndoNeRF [139]	✗	21.4	0.72	0.29	-	-	-
SAMSNeRF [75]	✔	**34.5**	**0.92**	**0.10**	-	-	-
DefSLAM [140]	✗	-	-	-	17.1	8.1	28.0
SuPerPM-F [76]	✔	-	-	-	**7.9**	**6.2**	**34.5**
Optical Flow Estimation						
Method	SAM-Based	Sintel		Sintel Occ.		KITTI-15
		clean	final	clean	final	F1
FlowFormer++ [6]	✗	1.07	**1.94**	6.64	10.63	4.52
SAMFlow [79]	✔	**1.00**	2.08	**5.97**	10.60	4.49
3D Point Cloud Segmentation						
Method	SAM-Based	ScanNet-2				
		wall	floor	cab	bed	Average
SparseConvNet [141]	✗	**83.2**	**94.8**	**61.9**	**76.9**	**68.2**
CSF [87]	✔	79.9	87.8	56.8	65.2	65.1

(Pre), and normalized precision ($nPre$)) [26], [128] are used. The one-pass evaluation (OPE), multi-start evaluation (MSE), and human-object interaction evaluation (HOI) are adopted on the TREK-150 [129] dataset. For the YouTube-VOS 2018 [123] dataset, the typical evaluation metrics include $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J} \& \mathcal{F}$ [119], [120].

Results Comparison. Tab. II showcases the performance of eleven representative SAM-based methods and six current SOTA trackers on VOTS2023, TREK-150, NUT-L, DAVIS 2017 test, DAVIS 2016 val, and YouTube-VOS 2018 val.

We make the following observations: **1)** DMAOT [128] achieves top performance with 63.6% AUC on the VOTS2023. This remarkable success can be attributed to the use of object-wise long term memory, which stores all masks of the tracked object in memory, and utilizes this memory to predict the current object mask, achieving more accurate results. **2)** Two SAM-based methods MSDeAOT [103] and SAM-DA [26] obtain the best results on the first person perspective tracking dataset TREK-150 and the long-term nighttime UAV tracking dataset NUT-L, respectively. The former deploys transformers at multiple feature scales and converts bounding boxes to refined masks with the help of SAM, while the latter utilizes a SAM-powered target domain training sample swelling strategy to dispose domain adaptation. From the results of MSDeAOT and SAM-DA, we can

observe that the powerful zero-shot generalization of SAM to generate precise masks and high-quality target domain training samples that are of great significance to specific video tasks. **3)** Using only click as initialization, the SAM-based method SAM-Track [27] scores 92.0% and 79.2% in terms of $\mathcal{J} \& \mathcal{F}$ on DAVIS 2016 val and DAVIS 2017 test, which are comparable to the current SOTA tracker SwinB-DeAOT-L [102]. **4)** SAM-based methods [16], [19], [69], [70] lag far behind the current SOTA tracker (STCN [132]) on the YouTube-VOS 2018 dataset. This is because many SAM-based methods lack the space time correspondences module. By modeling spatiotemporal correspondences in the context of video frames, STCN reduces memory usage and more

effectively utilizes information in memory and achieves top $\mathcal{J} \& \mathcal{F}$ with 83.0% on the YouTube-VOS 2018 val, which is significantly superior to its memory-based counterparts and SAM-based methods.

C. Evaluation of other Video Understanding Approaches

Datasets. FaceForensics++ [142] contains 1,004 videos, in which the forgery faces are generated by four deepfake algorithms (Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT)). ViSha [143] is a video shadow detection dataset, which comprises 120 videos, covering 7 shadow categories and 60 target categories with various lengths and different motion/lighting conditions. AVS-Bench [144] includes 4,932 videos, with two subsets: single source segmentation (V1S) and multiple sound source segmentation (V1M). Ref-DAVIS17 [145] and Ref-Youtube-VOS [146] are used to evaluate RVOS. Medical video datasets include EndoVis2017 [147], EndoVis2018 [148], CAMUS [149], In-house Data [5], EndoNeRF [139], and SuPer Data [76]. Sintel [150] and KITTI-15 [151] are common datasets for optical flow estimation; ScanNet-2 [152] is adopted in 3D point cloud segmentation.

Evaluation Metrics. In face forgery detection, accuracy is adopted for FaceForensics++ dataset [142]. For video shadow detection, the common metrics include mean absolute error (MAE), F-measure (F_β), intersection over union (IoU), and shadow/non-shadow region Balaence error rate (SBER/N-BER) [143]. Mean IoU (mIoU) and F-score are used on AVS-Bench [144]. Challenge IoU (cIoU), IoU, and mean class IoU (mcIoU) [74] are adopted on EndoVis2017 and EndoVis2018. For the CAMUS dataset, common evaluation metrics include Dice, dH, and dA [5]. For the EndoNeRF dataset, PSNR, SSIM, LPIPS [75] are used. The reprojection errors [76] is adopted to evaluate algorithms on the SuPer Data-V1, T1, and T2. We report the clean/final EPE [79] on Sintel test and Sintel test with occluded area, and F1 score on KITTI-15 [151]. The mIoU metric is used for the ScanNet-2 [152] dataset.

Results Comparison. Tab. III summarizes other video understanding results achieved by current SOTA and SAM-based methods. The main observations are presented as follows: **1)** Compared with VOS and VOT, SAM-based methods achieve more significant success in other video understanding tasks, *e.g.*, deepfake detection, video shadow detection, RVOS, and various medical videos. One common feature of these tasks is that they often have small-scale datasets. These results confirm that developing a foundation model typically pre-trained on a broad dataset is a promising direction to solve a range of downstream problems with limited data even distinct data distributions. **2)** For audio-visual segmentation, the carefully designed audio unmixing and semantic segmentation network (AUSS) [135] achieves top results on the single-source subset, but the SAM-based method GAVS [71] is more competitive on the complex multi-source subset. GAVS is a simple encoder-prompt-decoder paradigm leveraging the prior knowledge of the visual foundation model SAM. This highlights the importance of utilizing the generalization ability of foundation models to solve specific tasks. **3)** Compared

with fully supervised SparseConvNet [141], the SAM-based method CSF [87] achieves comparable results indicating that the effectiveness of adopting 2D foundation model to dispose 3D point cloud segmentation tasks. Inspired by the success of finetuning methods such as SuPerPM-F [76], we believe that finetuning foundation models on downstream tasks is a promising direction for addressing huge domain gaps.

D. Evaluation of Video Generation and Editing Approaches

Video Generation. We first summarize video synthesis results on Sequence of Pose Data [84] and TikTok Dance [85] datasets. The former contains 10 high-resolution human dance videos, while the later includes around 350 videos with video length of 10 to 15 seconds. For the Sequence of Pose Data, BRISQUE and NIQE [84] are introduced to assess the frame quality. We report frame-wise metrics (FID, SSIM, LISPIS, and PSNR), and video-wise metrics (FID-VID and FVD) [85] on the TikTok Dance dataset. Tab. IV presents the video generation quality evaluation of SAM-based methods (Dancing Avatar [84], DISCO [85]) and current SOTA methods (ControlVideo [153] and DreamPose [154]). It is evident that SAM-based methods exhibit a significant superiority compared to current SOTA methods in video generation tasks. Similar results can be observed from the VSR dataset Vimeo90K [114], and the SAM-based method (SEEM [25]) significantly surpasses the current SOTA method (BasicVSR [155]).

Video Editing. Tab. IV compares the results of generic video editing of the SAM-based method (Make-A-Protagonist [93]) and the current SOTA method (DreamBooth-V [156]), using model evaluation (CLIP [97] and DINO [158]) and user study (quality, subject, and prompt). Make-A-Protagonist achieves better model evaluation scores and an overwhelming preference in terms of user study. Furthermore, the SAM-based method (2SVE [94]) is significantly better than its counterpart (Text2Video-Zero [157]) on the challenging text guided video editing dataset TGVE [94].

Overall, the above results demonstrate that the visual foundation model SAM exhibits exceptional performance in video generation and editing tasks, although it is essentially for image segmentation. This recent trend of *SAM for anything*, has resulted in excellent performance in many other domains such as non-Euclidean domain [11] and adversarial attacks [29], paving the way to explore task-agnostic foundation models for vision and beyond [3].

VII. Conclusion and Future Directions

A. Conclusion

This survey offered an in-depth look at the latest developments in the era of foundation models with a focus on SAM for videos. To the best of our knowledge, this is the first systematic and comprehensive survey that concentrate on this specific and promising research field. We commenced by summarizing the unique challenges in the video domain, highlighting the extreme complexity of video tasks and the urgent need for a systematic review of SAM models for videos. This was followed by an overview of SAM and SAM 2,

TABLE IV: Performance evaluation of different video generation and video editing tasks. The best results are marked in **bold**.

Video Synthesis

Method	SAM-Based	Sequence of Pose Data [84]					
		Frame NIQE	Body NIQE	Background NIQE	Frame BRISQUE	Body BRISQUE	Background BRISQUE
ControlVideo [153]	✗	3.32	9.01	3.21	26.21	48.11	48.11
Dancing Avatar [84]	✔	**2.99**	**5.03**	**2.44**	**19.56**	**45.19**	**43.75**

Method	SAM-Based	TikTok Dance Dataset [85]					
		FID	SSIM	PSNR	LISPIS	FID-VID	FVD
DreamPose [154]	✗	72.62	0.511	28.11	0.442	78.77	551.02
DISCO [85]	✔	**28.31**	**0.674**	**29.15**	**0.285**	**55.17**	**267.75**

Video Super-Resolution

Method	SAM-Based	Vimeo-90K Fast [114]		Vimeo-90K Slow [114]		Vimeo-90K Average [114]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BasicVSR [155]	✗	38.2954	0.95152	32.5003	0.90629	35.3601	0.93287
SEEM [25]	✔	**38.4074**	**0.95222**	**32.6234**	**0.90775**	**35.4786**	**0.93390**

Video Editing

Method	SAM-Based	Model Evaluation		User Study		
		CLIP	DINO	Quality	Subject	Prompt
DreamBooth-V [156]	✗	0.301	**0.509**	22.81	24.69	15.00
Make-A-Protagonist [93]	✔	**0.329**	0.457	**67.50**	**63.44**	**66.25**

Method	SAM-Based	TGVE [94]					
		Text CLIPScore	Frame CLIPScore	PickScore	Text Alignment	Structure	Quality
Text2Video-Zero [157]	✗	25.88	**92.07**	19.82	0.448	**0.493**	0.516
2SVE [94]	✔	**26.89**	89.90	**20.71**	**0.689**	0.486	**0.599**

different research routes, and video-related research domains. Building upon above foundation, we exhaustively reviewed existing works and divided them into three key areas: video understanding, video generation, and video editing, considering their distinct technical perspectives and research objectives. In addition, we provided comparative results of SAM-based methods and current SOTA methods on various video tasks, together with numerous insightful observations.

B. Future Directions

Through our investigation and in-depth evaluation, we have found that although the SAM models (including SAM 2) has made or is making significant breakthroughs in various image and video tasks, there still exist numerous opportunities and challenges. We provide several future research directions in the area of SAM for videos and beyond in the following.

• **Constructing Large-Scale Video Datasets.** The substantial achievements of visual foundation models are mainly attributed to the availability of billions of high-quality image data. Nevertheless, considering the huge cost of data collection and annotation, current video tasks are usually limited to relatively small-scale datasets. For instance, the VOT dataset TrackingNet [159] contains 30,643 videos and 14.43 million frames, but its significant drawback is sparse annotation. Leveraging SAM to automatically generate dense mask annotations from videos is a potential solution to achieve data scalability [24].

• **Building Large-Scale Video Foundation Models.** Most current visual foundation models primarily concentrate on pre-training and adaptation at the image level, which are evidently constrained in complex and dynamic video-level understanding tasks. Due to the increasingly convenient collection and storage, videos are emerging as a domain force on the edge devices and Internet [22]. Therefore, the development of video foundation models, *e.g.*, medical video foundation models, for broad video applications becomes an urgent requirement.

• **Parameter-Efficient Training and Fast Inference.** Training video foundation models with billions of parameters from scratch inevitably faces significant challenges due to high data dimension and the high computational overhead. While some efforts to explore new technologies, *e.g.*, adapter [4] and prompt learning [160], by utilizing pre-trained models to promote efficient transfer learning, there remains a pressing need to mitigate training and inference expenses. More efficient training strategies and model compression methods may unlock more power in video foundation models on edge devices, *e.g.*, automobile and surgical robots, with limited computational resources.

• **Incorporating More Modalities.** Although current foundation models have achieved significant advances in single modality and two modalities (*e.g.*, vision and text, vision and audio), the integration of more modalities is far from being explored. A core reason is the lack of extensive aligned multi-modal data [2]. On one hand, collecting multi-modal data, *e.g.*, visual images, text, audio, point cloud, infrared images, depth images, and event streams, is crucial for researching multi-modal foundation models. On the other hand, developing a unified model [12] for multi-modal perception without requiring paired multi-modal data is a promising direction.

• **Credible and Interpretable Video Foundation Models.** The security of artificial intelligence has attracted significant concerns as it may lead to privacy breaches and security risks in practical applications such as face recognition and autonomous driving. However, the capability of video foundation models to resist various attacks [29] is still far from being explored. In addition, due to the high complexity and rapidly increasing deployment of video foundation models [3], improving their interpretability and enhancing people's trust in decision-making is a valuable avenue for future research.

• **More Innovative Opportunities in SAM for Videos.** As SAM for videos is a rapid-evolving research field, we might not cover all the latest advancements in this review. Actually,

there are masses of video tasks that are not covered by SAM or have not been fully studied, such as video captioning, video-based event detection, video-based behavior/action recognition, video summarization, and video frame interpolation. Last but not least, combining SAM with a large number of traditional technologies/methods (*e.g.*, knowledge distillation, and graph learning) and cutting-edge technologies/methods (*e.g.*, video diffusion model, explainable AI (XAI), and embodied AI) can stimulate more opportunities in the era of foundation models due to SAM's versatility and plug-and-play nature.

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 preprint arXiv:2108.07258*, 2021.

[2] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *arXiv preprint arXiv:2302.10035*, 2023.

[3] C. Zhang, L. Liu, Y. Cui, G. Huang, W. Lin, Y. Yang, and Y. Hu, "A comprehensive survey on segment anything model for vision and beyond," *arXiv preprint arXiv:2305.08196*, 2023.

[4] T. Chen, L. Zhu, C. Ding, R. Cao, S. Zhang, Y. Wang, Z. Li, L. Sun, P. Mao, and Y. Zang, "Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more," *arXiv preprint arXiv:2304.09148*, 2023.

[5] S. Kim, K. Kim, J. Hu, C. Chen, Z. Lyu, R. Hui, S. Kim, Z. Liu, A. Zhong, X. Li *et al.*, "Medivista-sam: Zero-shot medical video analysis with spatio-temporal sam adaptation," *arXiv preprint arXiv:2309.13539*, 2023.

[6] X. Shi, Z. Huang, D. Li, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li, "Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation," in *Computer Vision and Pattern Recognition*, 2023, pp. 1599–1610.

[7] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[8] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014, pp. 740–755.

[9] Y. Zhang and R. Jiao, "Towards segment anything model (sam) for medical image segmentation: A survey," *arXiv preprint arXiv:2305.03678*, 2023.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[11] Y. Jing, X. Wang, and D. Tao, "Segment anything in non-euclidean domains: Challenges and opportunities," *arXiv preprint arXiv:2304.11595*, 2023.

[12] Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue, "Meta-transformer: A unified framework for multimodal learning," *arXiv preprint arXiv:2307.10802*, 2023.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *International Conference on Computer Vision*, 2023, pp. 4015–4026.

[14] Y. Liu, M. Zhu, H. Li, H. Chen, X. Wang, and C. Shen, "Matcher: Segment anything with one shot using all-purpose feature matching," *arXiv preprint arXiv:2305.13310*, 2023.

[15] J. Fan, "Gpt-3 moment in computer vision," https://twitter.com/DrJimFan, 2023.

[16] L. Ke, M. Ye, M. Danelljan, Y. Liu, Y.-W. Tai, C.-K. Tang, and F. Yu, "Segment anything in high quality," *arXiv preprint arXiv:2306.01567*, 2023.

[17] S. Mo and Y. Tian, "Av-sam: Segment anything model meets audio-visual localization and segmentation," *arXiv preprint arXiv:2305.01836*, 2023.

[18] Y. Wang, W. Zhou, Y. Mao, and H. Li, "Detect any shadow: Segment anything for video shadow detection," *arXiv preprint arXiv:2305.16698*, 2023.

[19] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, P. Gao, and H. Li, "Personalize segment anything model with one shot," *arXiv preprint arXiv:2305.03048*, 2023.

[20] 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 preprint arXiv:2305.12659*, 2023.

[21] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint*, 2024.

[22] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, "A survey on video diffusion models," *arXiv preprint arXiv:2310.10647*, 2023.

[23] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," *arXiv preprint arXiv:2304.11968*, 2023.

[24] H. He, J. Zhang, M. Xu, J. Liu, B. Du, and D. Tao, "Scalable mask annotation for video text spotting," *arXiv preprint arXiv:2305.01443*, 2023.

[25] Z. Lu, Z. Xiao, J. Bai, Z. Xiong, and X. Wang, "Can sam boost video super-resolution?" in *arXiv preprint arXiv:2305.06524*, 2023.

[26] L. Yao, H. Zuo, G. Zheng, C. Fu, and J. Pan, "Sam-da: Uav tracks anything at night with sam-powered domain adaptation," *arXiv preprint arXiv:2307.01024*, 2023.

[27] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," *arXiv preprint arXiv:2305.06558*, 2023.

[28] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," *arXiv preprint arXiv:2306.12156*, 2023.

[29] C. Zhang, C. Zhang, T. Kang, D. Kim, S.-H. Bae, and I. S. Kweon, "Attack-sam: Towards evaluating adversarial robustness of segment anything model," *arXiv preprint arXiv:2305.00866*, 2023.

[30] Y. Lai, Z. Luo, and Z. Yu, "Detect any deepfakes: Segment anything meets face forgery detection and localization," *arXiv preprint arXiv:2306.17075*, 2023.

[31] C. Zhang, S. Zheng, C. Li, Y. Qiao, T. Kang, X. Shan, C. Zhang, C. Qin, F. Rameau, S.-H. Bae *et al.*, "A survey on segment anything model (sam): Vision foundation model meets prompt engineering," *arXiv preprint arXiv:2306.06211*, 2023.

[32] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz *et al.*, "Augmented language models: a survey," *arXiv preprint arXiv:2302.07842*, 2023.

[33] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," *arXiv preprint arXiv:2304.02020*, 2023.

[34] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of" bias" in nlp," *arXiv preprint arXiv:2005.14050*, 2020.

[35] S. Long, F. Cao, S. C. Han, and H. Yang, "Vision-and-language pretrained models: A survey," *arXiv preprint arXiv:2204.07356*, 2022.

[36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[37] 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," in *International Conference on Learning Representations*, 2021.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *International Conference on Computer Vision*, 2017, pp. 2980–2988.

[39] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 2016, pp. 565–571.

[40] Y. Zhao *et al.*, "Gait-assisted video person retrieval," *IEEE TCSVT*, vol. 33, no. 2, pp. 897–908, 2023.

[41] Z. Zheng *et al.*, "Dynamic spatial focus for efficient compressed video action recognition," *IEEE TCSVT*, pp. 1–1, 2023.

[42] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[43] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *International Conference on Computer Vision*, 2019, pp. 6202–6211.

[44] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[45] A. Stoian *et al.*, "Fast action localization in large-scale video archives," *IEEE TCSVT*, vol. 26, no. 10, pp. 1917–1930, 2016.

[46] J. Tan, J. Tang, L. Wang, and G. Wu, "Relaxed transformer decoders for direct action proposal generation," in *International Conference on Computer Vision*, 2021, pp. 13 526–13 535.

[47] M. Nawhal and G. Mori, "Activity graph transformer for temporal action localization," *arXiv preprint arXiv:2101.08540*, 2021.

[48] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *European Conference on Computer Vision*, 2022, pp. 492–510.

[49] X. Tu *et al.*, "Image-to-video generation via 3d facial dynamics," *IEEE TCSVT*, vol. 32, no. 4, pp. 1805–1819, 2022.

[50] M. Siu *et al.*, "A robust model generation technique for model-based video coding," *IEEE TCSVT*, vol. 11, no. 11, pp. 1188–1192, 2001.

[51] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[52] 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 *European conference on computer vision*, 2022, pp. 720–736.

[53] 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 preprint arXiv:2210.02303*, 2022.

[54] 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 preprint arXiv:2209.14792*, 2022.

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

[56] R. Arora and Y. J. Lee, "Singan-gif: Learning a generative video model from a single gif," in *Winter Conference on Applications of Computer Vision*, 2021, pp. 1310–1319.

[57] S. Gur, S. Benaim, and L. Wolf, "Hierarchical patch vae-gan: Generating diverse videos from a single sample," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 761–16 772, 2020.

[58] N. Haim, B. Feinstein, N. Granot, A. Shocher, S. Bagon, T. Dekel, and M. Irani, "Diverse generation from a single video made possible," in *European Conference on Computer Vision*, 2022, pp. 491–509.

[59] Y. Nikankin, N. Haim, and M. Irani, "Sinfusion: Training diffusion models on a single image or video," *arXiv preprint arXiv:2211.11743*, 2022.

[60] O. Jamriška, Š. Sochorová, O. Texler, M. Lukáč, J. Fišer, J. Lu, E. Shechtman, and D. Sýkora, "Stylizing video by example," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–11, 2019.

[61] O. Texler, D. Futschik, M. Kučera, O. Jamriška, Š. Sochorová, M. Chai, S. Tulyakov, and D. Sýkora, "Interactive video stylization using few-shot patch-based training," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 73–1, 2020.

[62] Y. Kasten, D. Ofri, O. Wang, and T. Dekel, "Layered neural atlases for consistent video editing," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–12, 2021.

[63] E. Lu, F. Cole, T. Dekel, A. Zisserman, W. T. Freeman, and M. Rubinstein, "Omnimatte: Associating objects and their effects in video," in *Computer Vision and Pattern Recognition*, 2021, pp. 4507–4515.

[64] S. Chang, Z. Hao, B. Kang, X. Zhao, J. Zhu, Z. Chen, L. Zhang, L. Zhang, and H. Lu, "3rd place solution for pvuw2023 vss track: A large model for semantic segmentation on vspw," *arXiv preprint arXiv:2306.02291*, 2023.

[65] Z. Zhou, Z. Wu, R. Boutteau, F. Yang, and D. Ginhac, "Dsec-mos: Segment any moving object with moving ego vehicle," *arXiv preprint arXiv:2305.00126*, 2023.

[66] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee, "Tracking anything with decoupled video segmentation," in *International Conference on Computer Vision*, 2023, pp. 1316–1326.

[67] L. Qi, J. Kuen, T. Shen, J. Gu, W. Li, W. Guo, J. Jia, Z. Lin, and M.-H. Yang, "High-quality entity segmentation," in *International Conference on Computer Vision*, 2023, pp. 4047–4056.

[68] J. Zhu, Z. Chen, Z. Hao, S. Chang, L. Zhang, D. Wang, H. Lu, B. Luo, J.-Y. He, J.-P. Lan *et al.*, "Tracking anything in high quality," *arXiv preprint arXiv:2307.13974*, 2023.

[69] W.-H. Chu, A. W. Harley, P. Tokmakov, A. Dave, L. Guibas, and K. Fragkiadaki, "Zero-shot open-vocabulary tracking with large pre-trained models," *arXiv preprint arXiv:2310.06992*, 2023.

[70] F. Rajič, L. Ke, Y.-W. Tai, C.-K. Tang, M. Danelljan, and F. Yu, "Segment anything meets point tracking," *arXiv preprint arXiv:2307.01197*, 2023.

[71] Y. Wang, W. Liu, G. Li, J. Ding, D. Hu, and X. Li, "Prompting segmentation with sound is generalizable audio-visual source localizer," *arXiv preprint arXiv:2309.07929*, 2023.

[72] S. Bhosale, H. Yang, D. Kanojia, and X. Zhu, "Leveraging foundation models for unsupervised audio-visual segmentation," *arXiv preprint arXiv:2309.06728*, 2023.

[73] Y. Li, J. Zhang, X. Teng, and L. Lan, "Refsam: Efficiently adapting segmenting anything model for referring video object segmentation," *arXiv preprint arXiv:2307.00997*, 2023.

[74] W. Yue, J. Zhang, K. Hu, Y. Xia, J. Luo, and Z. Wang, "Surgicalsam: Efficient class promptable surgical instrument segmentation," *arXiv preprint arXiv:2308.08746*, 2023.

[75] A. Lou, Y. Li, X. Yao, Y. Zhang, and J. Noble, "Samsnerf: Segment anything model (sam) guides dynamic surgical scene reconstruction by neural radiance field (nerf)," *arXiv preprint arXiv:2308.11774*, 2023.

[76] S. Lin, A. J. Miao, A. Alabiad, F. Liu, K. Wang, J. Lu, F. Richter, and M. C. Yip, "Superpm: A large deformation-robust surgical perception framework based on deep point matching learned from physical constrained simulation data," *arXiv preprint arXiv:2309.13863*, 2023.

[77] M. E. Bonani, M. Schwarz, and S. Behnke, "Learning from sam: Harnessing a segmentation foundation model for sim2real domain adaptation through regularization," *arXiv preprint arXiv:2309.15562*, 2023.

[78] C.-Y. Hsieh, S.-A. Chen, C.-L. Li, Y. Fujii, A. Ratner, C.-Y. Lee, R. Krishna, and T. Pfister, "Tool documentation enables zero-shot tool-usage with large language models," *arXiv preprint arXiv:2308.00675*, 2023.

[79] S. Zhou, R. He, W. Tan, and B. Yan, "Samflow: Eliminating any fragmentation in optical flow with segment anything model," *arXiv preprint arXiv:2307.16586*, 2023.

[80] J. Yang, W. Tan, C. Jin, B. Liu, J. Fu, R. Song, and L. Wang, "Pave the way to grasp anything: Transferring foundation models for universal pick-place robots," *arXiv preprint arXiv:2306.05716*, 2023.

[81] A. D. Raha, M. S. Munir, A. Adhikary, Y. Qiao, and C. S. Hong, "Generative ai-driven semantic communication framework for nextg wireless network," *arXiv preprint arXiv:2310.09021*, 2023.

[82] C. A. Schiller, "Virtual augmented reality for atari reinforcement learning," *arXiv preprint arXiv:2310.08683*, 2023.

[83] B. Benjdira, A. Koubaa, and A. M. Ali, "Rosgpt_vision: Commanding robots using only language models' prompts," *arXiv preprint arXiv:2308.11236*, 2023.

[84] B. Qin, W. Ye, Q. Yu, S. Tang, and Y. Zhuang, "Dancing avatar: Pose and text-guided human motion videos synthesis with image diffusion model," *arXiv preprint arXiv:2308.07749*, 2023.

[85] T. Wang, L. Li, K. Lin, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Disco: Disentangled control for realistic human dance generation," *arXiv preprint arXiv:2307.00040*, 2023.

[86] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, "Sam3d: Segment anything in 3d scenes," *arXiv preprint arXiv:2306.03908*, 2023.

[87] S. Dong, F. Liu, and G. Lin, "Leveraging large-scale pretrained vision foundation models for label-efficient 3d point cloud segmentation," *arXiv preprint arXiv:2311.01989*, 2023.

[88] J. Xu, W. Zhao, Z. Tang, and X. Gan, "A one stop 3d target reconstruction and multilevel segmentation method," *arXiv preprint arXiv:2308.06974*, 2023.

[89] D. Balaban, J. Medich, P. Gosar, and J. Hart, "Propagating semantic labels in video data," *arXiv preprint arXiv:2310.00783*, 2023.

[90] R. Guo, Y. Chen, Y. Qi, W. Yue, D. Niu, and X. Ying, "Audio-visual instance segmentation," *arXiv preprint arXiv:2310.18709*, 2023.

[91] T. Delatolas, V. Kalogeiton, and D. P. Papadopoulos, "Learning the what and how of annotation in video object segmentation," *WACV*, 2023.

[92] B. Kiefer, T. Höfer, and A. Zell, "Stable yaw estimation of boats from the viewpoint of uavs and usvs," *arXiv preprint arXiv:2306.14056*, 2023.

[93] Y. Zhao, E. Xie, L. Hong, Z. Li, and G. H. Lee, "Make-a-protagonist: Generic video editing with an ensemble of experts," *arXiv preprint arXiv:2305.08850*, 2023.

[94] J. Z. Wu, X. Li, D. Gao, Z. Dong, J. Bai, A. Singh, X. Xiang, Y. Li, Z. Huang, Y. Sun *et al.*, "Cvpr 2023 text guided video editing competition," *arXiv preprint arXiv:2310.16003*, 2023.

[95] Y. Yin, Z. Fu, F. Yang, and G. Lin, "Or-nerf: Object removing from 3d scenes guided by multiview segmentation with neural radiance fields," *arXiv preprint arXiv:2305.10503*, 2023.

[96] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai, "In defense of online models for video instance segmentation," in *European Conference on Computer Vision*, 2022, pp. 588–605.

[97] 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 *International conference on machine learning*, 2021, pp. 8748–8763.

[98] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[99] J. Miao, X. Wang, Y. Wu, W. Li, X. Zhang, Y. Wei, and Y. Yang, "Large-scale video panoptic segmentation in the wild: A benchmark," in *Computer Vision and Pattern Recognition*, 2022, pp. 21 033–21 043.

[100] X. Li, W. Zhang, J. Pang, K. Chen, G. Cheng, Y. Tong, and C. C. Loy, "Video k-net: A simple, strong, and unified baseline for video segmentation," in *Computer Vision and Pattern Recognition*, 2022, pp. 18 847–18 857.

[101] H. K. Cheng and A. G. Schwing, "Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *European Conference on Computer Vision*, 2022, pp. 640–658.

[102] Z. Yang and Y. Yang, "Decoupling features in hierarchical propagation for video object segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 324–36 336, 2022.

[103] Y. Xu, J. Li, Z. Yang, Y. Yang, and Y. Zhuang, "Zju reler submission for epic-kitchen challenge 2023: Trek-150 single object tracking," *arXiv preprint arXiv:2307.02508*, 2023.

[104] X. Zhou, C. Yang, H. Zhao, and W. Yu, "Low-rank modeling and its applications in image analysis," *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, pp. 1–33, 2014.

[105] Z. Qiu, Y. Hu, H. Li, and J. Liu, "Learnable ophthalmology sam," *arXiv preprint arXiv:2304.13425*, 2023.

[106] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 10, p. 1337–1342, 2003.

[107] A. Ecins, C. Fermˋuller, and Y. Aloimonos, "Shadow free segmentation in still images using local density measure," in *2014 IEEE International Conference on Computational Photography (ICCP)*, 2014, pp. 1–8.

[108] D. Liu, C. Long, H. Zhang, H. Yu, X. Dong, and C. Xiao, "Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes," in *Computer Vision and Pattern Recognition*, 2020, p. 8139–8148.

[109] X. Xu, J. Wang, X. Li, and Y. Lu, "Reliable propagation-correction modulation for video object segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, p. 2946–2954.

[110] A. Wang, M. Islam, M. Xu, Y. Zhang, and H. Ren, "Sam meets robotic surgery: An empirical study in robustness perspective," *arXiv preprint arXiv:2304.14674*, 2023.

[111] L. Fillioux, E. Gontran, J. Cartry, J. R. Mathieu, S. Bedja, A. Boilève, P.-H. Cournède, F. Jaulin, S. Christodoulidis, and M. Vakalopoulou, "Spatio-temporal analysis of patient-derived organoid videos using deep learning for the prediction of drug efficacy," in *International Conference on Computer Vision*, 2023, pp. 3930–3939.

[112] T. Ren, S. Liu, A. Zeng, H. Cao, J. Lin, F. Li, H. Zhang, H. Li, Z. Zeng, and L. Zhang, "Grounded-sam: Detect and segment anything with text prompt," *ICCV Demo*, 2023.

[113] A. Kannan, "Learning from human videos for robotic manipulation," Ph.D. dissertation, Carnegie Mellon University Pittsburgh, PA, 2023.

[114] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.

[115] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv preprint arXiv: 2201.03546*, 2022.

[116] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller *et al.*, "Openclip," *Zenodo*, vol. 4, p. 5, 2021.

[117] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *International Conference on Computer Vision*, 2023, pp. 3836–3847.

[118] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 190–200. [Online]. Available: https://aclanthology.org/P11-1020

[119] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017.

[120] 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 *Computer Vision and Pattern Recognition*, 2016, pp. 724–732.

[121] W. Wang, M. Feiszli, H. Wang, and D. Tran, "Unidentified video objects: A benchmark for dense, open-world segmentation," in *International Conference on Computer Vision*, 2021, pp. 10 776–10 785.

[122] L. Ke, H. Ding, M. Danelljan, Y.-W. Tai, C.-K. Tang, and F. Yu, "Video mask transfiner for high-quality video instance segmentation," in *European Conference on Computer Vision*, 2022, pp. 731–747.

[123] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "Youtube-vos: Sequence-to-sequence video object segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 585–601.

[124] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[125] Y. Du, W. Guo, Y. Xiao, and V. Lepetit, "Uvo challenge on video-based open-world segmentation 2021: 1st place solution," *arXiv preprint arXiv:2110.11661*, 2021.

[126] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit *et al.*, "Openimages: A public dataset for large-scale multi-label and multi-class image classification," *Dataset available from https://github. com/openimages*, vol. 2, no. 3, p. 18, 2017.

[127] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[128] M. Kristan, J. Matas, M. Danelljan, M. Felsberg, H. J. Chang, L. Č. Zajc, A. Lukežič, O. Drbohlav, Z. Zhang, K.-T. Tran *et al.*, "The first visual object tracking segmentation vots2023 challenge results," in *International Conference on Computer Vision*, 2023, pp. 1796–1818.

[129] M. Dunnhofer, A. Furnari, G. M. Farinella, and C. Micheloni, "Visual object tracking in first person vision," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 259–283, 2023.

[130] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, "High-performance long-term tracking with meta-updater," in *Computer Vision and Pattern Recognition*, 2020, pp. 6298–6307.

[131] J. Ye, C. Fu, G. Zheng, D. P. Paudel, and G. Chen, "Unsupervised domain adaptation for nighttime aerial tracking," in *Computer Vision and Pattern Recognition*, 2022, pp. 8896–8905.

[132] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 781–11 794, 2021.

[133] C. Kong, B. Chen, H. Li, S. Wang, A. Rocha, and S. Kwong, "Detect and locate: Exposing face manipulation by semantic-and noise-level telltales," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1741–1756, 2022.

[134] L. Liu, J. Prost, L. Zhu, N. Papadakis, P. Liò, C.-B. Schönlieb, and A. I. Aviles-Rivero, "Scotch and soda: A transformer video shadow detection framework," in *Computer Vision and Pattern Recognition*, 2023, pp. 10 449–10 458.

[135] Y. Ling, Y. Li, Z. Gan, J. Zhang, M. Chi, and Y. Wang, "Hear to segment: Unmixing the audio to guide the semantic segmentation," *arXiv preprint arXiv:2305.07223*, 2023.

[136] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo, "Language as queries for referring video object segmentation," in *Computer Vision and Pattern Recognition*, 2022, pp. 4974–4984.

[137] N. Ayobi, A. Pérez-Rondón, S. Rodríguez, and P. Arbeláez, "Matis: Masked-attention transformers for surgical instrument segmentation," *arXiv preprint arXiv:2303.09514*, 2023.

[138] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain

tumors in mri images," in *International MICCAI Brainlesion Workshop*, 2021, pp. 272–284.

[139] Y. Wang, Y. Long, S. H. Fan, and Q. Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 431–441.

[140] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, "Defslam: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Transactions on robotics*, vol. 37, no. 1, pp. 291–303, 2020.

[141] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Computer Vision and Pattern Recognition*, 2018, pp. 9224–9232.

[142] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.

[143] Z. Chen, L. Wan, L. Zhu, J. Shen, H. Fu, W. Liu, and J. Qin, "Triple-cooperative video shadow detection," in *Computer Vision and Pattern Recognition*, 2021, pp. 2715–2724.

[144] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, "Audio–visual segmentation," in *European Conference on Computer Vision*, 2022, pp. 386–403.

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

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

[147] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[148] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamoham-madi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen *et al.*, "2018 robotic scene segmentation challenge," *arXiv preprint arXiv:2001.11190*, 2020.

[149] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[150] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, 2012, pp. 611–625.

[151] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[152] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.

[153] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "Controlvideo: Training-free controllable text-to-video generation," *arXiv preprint arXiv:2305.13077*, 2023.

[154] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "Dreampose: Fashion image-to-video synthesis via stable diffusion," *arXiv preprint arXiv:2304.06025*, 2023.

[155] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," in *Computer Vision and Pattern Recognition*, 2021, pp. 4947–4956.

[156] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *International Conference on Computer Vision*, 2023, pp. 7623–7633.

[157] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," *arXiv preprint arXiv:2303.13439*, 2023.

[158] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *International Conference on Computer Vision*, 2021, pp. 9650–9660.

[159] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 300–317.

[160] C. Li, P. Khanduri, Y. Qiang, R. I. Sultan, I. Chetty, and D. Zhu, "Auto-prompting sam for mobile friendly 3d medical image segmentation," *arXiv preprint arXiv:2308.14936*, 2023.