OmniViD: A Generative Framework for Universal Video Understanding

Junke Wang[1,2], Dongdong Chen[3], Chong Luo[4], Bo He[5], Lu Yuan[3], Zuxuan Wu[1,2†], Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3]Microsoft Cloud + AI, [4]Microsoft Research Asia, [5]University of Maryland, College Park

Abstract

The core of video understanding tasks, such as recognition, captioning, and tracking, is to automatically detect objects or actions in a video and analyze their temporal evolution. Despite sharing a common goal, different tasks often rely on distinct model architectures and annotation formats. In contrast, natural language processing benefits from a unified output space, i.e., text sequences, which simplifies the training of powerful foundational language models, such as GPT-3, with extensive training corpora. Inspired by this, we seek to unify the output space of video understanding tasks by using languages as labels and additionally introducing time and box tokens. In this way, a variety of video tasks could be formulated as video-grounded token generation. This enables us to address various types of video tasks, including classification (such as action recognition), captioning (covering clip captioning, video question answering, and dense video captioning), and localization tasks (such as visual object tracking) within a fully shared encoder-decoder architecture, following a generative framework. Through comprehensive experiments, we demonstrate such a simple and straightforward idea is quite effective and can achieve state-of-the-art or competitive results on seven video benchmarks, providing a novel perspective for more universal video understanding. Code is available at https://github.com/wangjk666/OmniVid.

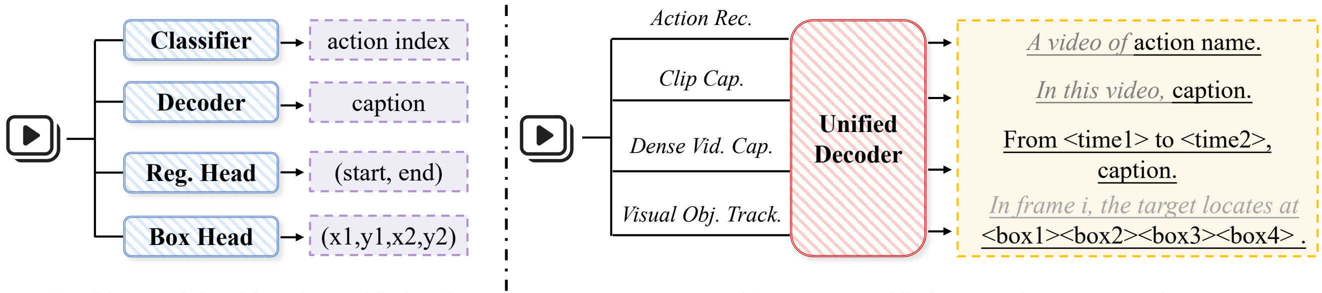
1. Introduction

In recent years, the proliferation of video content across various applications, such as online education and live streaming, has profoundly impacted our daily lives. Videos have evolved into a captivating and immersive medium for information delivery, emphasizing the pressing demand for the development of automated algorithms capable of understanding the actions [46], events [49], and moving objects [81] within video sequences. As a result, the field of video understanding has undergone significant expansion and encompassed a diverse range of tasks, including action recognition [3, 33, 60, 64, 80, 103], video captioning [16, 36, 61], and object tracking [4, 20, 95, 116].

For a long period, research in video understanding has often adopted a task-specific paradigm, *i.e.*, designing specialized architectures and loss functions to cater to the unique requirements of different tasks and benchmarks [10, 46, 69, 107]. Despite the promising results with high-capacity deep neural networks, these methods [30, 82, 104, 113] are tailored for a particular objective and less adaptable to deployment in scenarios of diverse needs. To mitigate this issue, video foundation models [92, 93, 101, 105], have gained emerging attention for their impressive performance across a broad spectrum of video tasks and potential in realizing the vision of Artificial General Intelligence (AGI). However, while generic spatial-temporal representations can be learned with these models, adapting them to different downstream tasks oftentimes requires carefully designing and fine-tuning task-specific heads.

In this paper, we posit such limitation originates from the diversified annotations for different video tasks, *e.g.*, a set of action categories for action recognition [12, 80, 103], sentences for captioning [36, 61], and continuous segments (coordinates) for events (object) localization [20, 23, 71]. This naturally necessitates task-specific designs for better optimization. In contrast, different tasks in natural language processing (NLP) enjoy a sharable output space, *i.e.*, text sequences, which promotes the development of large language models, such as GPT [77, 78] and Llama [45, 83, 84]. Drawing inspiration from this, we leverage word tokens in natural languages to represent semantic information that is important for coarse-grained tasks like action recognition, video captioning, and video question answering, and additionally introduce special *time tokens* and *box tokens* that provide localization capabilities in both spatial and temporal dimensions, particularly useful for fine-grained tasks like dense video captioning and visual object tracking. With such an enriched vocabulary that consists of word, time, and box tokens, the output format, as well as training objectives

[†]Corresponding author.



(a) Video models with task-specific heads

(b) Ours - A Unified Generative Framework

Figure 1. A conceptual comparison between existing video models and OmniViD.

of different tasks, can be well unified. Please refer to Figure 1 for a better illustration.

With this in mind, we present OmniViD, a generative framework that approaches various video tasks as a language modeling task conditioned on video inputs. OmniViD adopts an encoder-decoder architecture, where a dedicated video encoder and a language encoder are employed to extract the multimodal features from diverse inputs. Considering the remarkable redundancy in video data, we propose a lightweight MQ-former to enhance the efficiency of video representations for subsequent modeling. The MQ-former utilizes three types of learnable queries, *i.e*., content, sentence, and box queries, to aggregate the frame features from the video encoder through cross-attention. Finally, a token decoder is applied to generate a token sequence from the above vocabulary.

We validate the effectiveness of OmniViD on five representative video tasks, including action recognition, clip captioning, video question answering, dense video captioning, and visual object tracking. The results demonstrate that OmniViD achieves new state-of-the-art or at least competitive results on the prevalent video benchmarks. For example, using VideoSwin-Base [64] as the video encoder, we achieve state-of-the-art performance on action recognition (83.6% top1 accuracy on Kinetics-400 [46]), clip captioning (56.6 on MSRVTT [107] in terms of CIDEr), video question answering (42.3% accuracy on MSRVTT [107]), dense video captioning (5.6 on ActivityNet [10] in terms of SODA$_c$), and visual object tracking (88.9 on TrackingNet [69] in terms of normalized precision). For the first time, video tasks of different modalities and granularity can be supported by a single framework.

2. Related Work

2.1. Task-specific Methods for Video Understanding

Task-specific video understanding models could be roughly divided into classification, captioning, and localization approaches. Video action recognition is the most representative classification task in the video domain, which aims to recognize human actions in a video. Existing methods, including both CNN-based [32, 33, 46, 66] and Transformer-based models [3, 30, 64], widely encode the action labels as one-hot vectors and employ cross-entropy loss for supervised training. Captioning tasks, on the other hand, typically generate a textual description for a video clip [61, 125, 126] or an untrimmed long video [44, 99, 113] with a text decoder like BERT [47]. It is worth noting that captioning long videos involves the additional challenge of temporal event localization within the video, making it a more complex task. We categorize the open-ended video question answering [50, 51, 59] as a specific type of captioning task due to the consistent output format between them. Localization tasks, represented by visual object tracking [20, 25, 96], estimate the trajectory of a target object in a video sequence given its position in the first frame. Following the practice in object detection [11, 38, 40], a box head is oftentimes adopted to regress the coordinates of the tracking object. In summary, divergent prediction heads have been developed in various video tasks to adapt to the specific format of annotations, which poses a challenge to derive a unified solution. In this paper, we rethink the design of a universal video understanding framework from a novel perspective, *i.e*., redefining an output space that could be shared by different video tasks. Within this unified space, the development of general architectures and training objectives become distinctly feasible.

2.2. Unified Video Models

Recently, researchers have undertaken prominent efforts to unify video tasks within specific domains. OmniVL [92] and InterVideo [101] represent significant strides in the realm of video-language pretraining, which are pre-trained on large-scale video-text data and achieve superior results on multimodal video tasks like text-to-video retrieval and video captioning. Beyond these advancements, UNLoc [111] and UniVTG [76] have sought to tackle a diverse array of temporal localization tasks within a single framework. They accomplish this by simultaneously predicting saliency scores and boundary offsets for each frame (clip). Compared to video-language and temporal localization, spatial localization in the video domain, *i.e*., tracking,

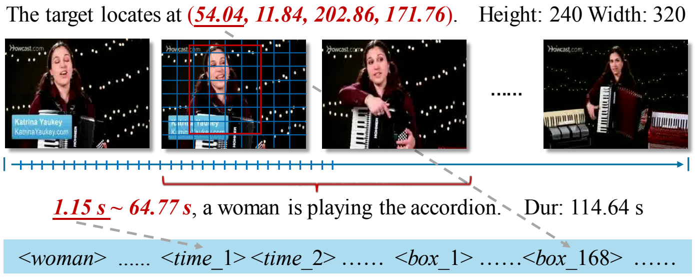

The target locates at (54.04, 11.84, 202.86, 171.76). Height: 240 Width: 320
......
1.15 s ~ 64.77 s, a woman is playing the accordion. Dur: 114.64 s
<woman> ...... <time_1> <time_2> ...... <box_1> ......<box_168> ......

Figure 2. Illustration of the *time tokens* and *box tokens* in OmniViD.

Table 1. Input & output of different video tasks. S/B/W/T: Sentence / Box / Word / Time, Pro. / Tok.: Prompt / Token.

Task	**Input**		**Target**		
	S Pro.	B Pro.	W Tok.	T Tok.	B Tok.
AR	✗	✗	✓	✗	✗
CC	✗	✗	✓	✗	✗
ViQA	✓	✗	✓	✗	✗
DVP	✗	✗	✓	✓	✗
VOT	✗	✓	✗	✗	✓

is more fragmented in terms of task definition, model architecture, and benchmarks. Unicorn [109] marks a significant step forward by employing a fully shared CNN-based encoder and box head for various tracking tasks, utilizing a target before distinguishing between them. Subsequently, with the prominent success of vision transformer [11], OmniTracker [94] and UNINEXT [110] push the boundaries of unification in tracking models by incorporating Transformer-based detectors. Despite the achievements of these approaches, they are still constrained by task-specific heads, leaving considerable space for greater unification of video understanding. To address this limitation, we unify diverse tasks with a sharable output space and address them with a fully shared generative framework.

2.3. Autoregress Modeling in Computer Vision

AutoRegressive modeling [114] is a statistical modeling technique that predicts the current state of a sequence based on historical observations, which has achieved remarkable success in the field of natural language processing (NLP) [24] and time series analysis [34, 68]. Inspired by this, researchers in the vision community have also attempted to explore its potential for visual understanding. Pix2SeqV1&V2 [18, 19] expand the textual vocabulary with quantized image coordinates. With this, they address several fundamental image tasks, *e.g.*, object detection, and image captioning, in a unified autoregressive manner. Following this idea, ARTrack [102] and SeqTrack [21] further support the visual object tracking task. VisionLLM [100], on the other hand, directly builds vision-centric frameworks upon pre-trained LLMs, with the hope of transferring their knowledge to visual understanding with minimal resource overhead. In this work, we leverage autoregressive modeling to the design of a universal video understanding framework. In addition to the expansion to temporal localization tasks with unique *time tokens*, our method also explores the advantages of autoregressive modeling for a universal video understanding framework for the first time.

3. Method

Our primary objective is to design a universal framework that accommodates a diverse set of video understanding tasks. To accomplish this, we expand upon the vocabulary commonly used in language models [8, 53] by introducing unique *time tokens* and *box tokens*. This augmentation allows us to represent the output of various video tasks as a token sequence within a shared vocabulary. Building upon this foundation, we further present OmniViD, a generative framework that conceptualizes video tasks as a process for generating tokens grounded in the video content.

Given a video $\mathcal{V}$ that lasts tens of seconds to multiple minutes, we sample a sequence of frames $[X_1, X_2, ..., X_T]$ from it. For video question answering, a question regarding the visual content is given, while for visual object tracking, the bounding box of the target object in the first frame is specified by the user. Below we first introduce how to perform tokenization for different video tasks with the above vocabulary in Sec. 3.1, and then present the architecture of OmniViD in Sec. 3.2. Finally, we elaborate on the unified training and inference pipeline in Sec. 3.3.

3.1. Unified Vocabulary for Video Understanding

In video understanding, various tasks necessitate diverse inputs and outputs according to specific settings and requirements. To establish a cohesive output space that could be shared by different video tasks, we supplement the word tokens in language vocabulary with special *time tokens* and *box tokens*, by discretizing the timestamps and the coordinates along the temporal and spatial dimensions, respectively (see Figure 2).

With the enriched vocabulary, the input and target sequences for the training of OmniViD can be generated in the following manner:

- **Action Recognition**: the input only includes a task prompt p_{task}, *i.e.*, “action recognition”, and the target is the ground-truth action name, *e.g.*, “dancing ballet”.
- **Clip Captioning**: similar to action recognition, the only difference lies in the target sequence becomes a longer description, *e.g.*, “a clip showing a computer screen”.
- **Video Question-Answering**: the input includes both the task prompt and the question p_{sen}, *e.g.*, “What is the video doing?”, while the target is the answer to that question, *e.g.*, “fencing competition”.

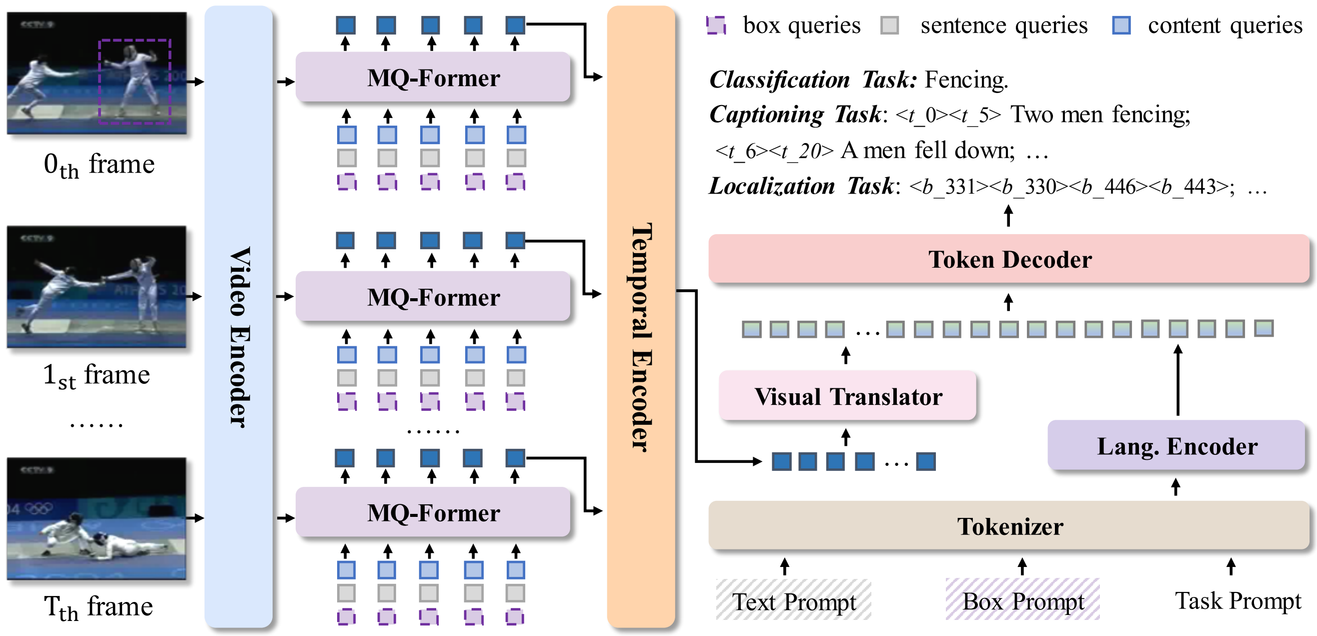


Figure 3. Architecture of OmniViD. The Mixed Q-former aggregates the frame features into three types of queries, *i.e.*, content queries, text queries, and box queries. After that, the queries obtained from different frames are input to a temporal encoder for temporal modeling. Finally, the token decoder generates a sequence of tokens conditioned on the multimodal inputs.

- **Dense Video Captioning**: the expected output is a set of events $\{e_i\}_{i=1}^{E}$ happening in the given video. In order to facilitate the model to learn the correspondence between timestamps and visual contents, we define a triplet for the i-th event e_i: $e_i = \langle t_i^{start}, t_i^{dur}, s\rangle$, where t_i^{start} and t_i^{dur} denote the start and duration time token, and s represents the description for the event [113]. The target sequence is constructed by concatenating the triplets of all the events.
- **Visual Object Tracking**: we take the task prompt and the discrete representation of the bounding box in the first frame, p_{box}, as input, and employ the *box tokens* in the following frames as target. Given a bounding box (x1, y1, x2, y2) on an $H \times W$ image, the tokenized representation is $(\langle box\_\lfloor \mathrm{x1}/W\rfloor\rangle$, $\langle box\_\lfloor \mathrm{y1}/H\rfloor\rangle$, $\langle box\_\lfloor \mathrm{x2}/W\rfloor\rangle$, $\langle box\_\lfloor \mathrm{y2}/H\rfloor\rangle)$.

The input and target sequence for different video tasks are summarized in Table 1.

3.2. Unified Architecture

OmniViD follows an encoder-decoder architecture, which first extracts the video features $F \in \mathcal{R}^{T^f \times H^f \times W^f \times C^f}$ from $\{X_t\}_{t=1}^{T}$ with a video encoder, where T^f and $H^f \times W^f$ denote the temporal and spatial resolution and C^f is the feature dimension. For visual object tracking, we replace the first frame with the cropped template, following the common practice [4, 20, 96]. A language encoder is also adopted to transform three types of prompts, p_{task}, p_{sen}, p_{box} to the prompt embeddings G_{task}, G_{sen}, G_{box}, and then concatenate them as the textual feature $G \in \mathcal{R}^{L^g \times C^g}$ along the sequence dimension. Based on the multimodal inputs, OmniViD produces a sequence of tokens in the above vocabulary. The overall framework is illustrated in Figure 3.

MQ-former. In order to encode the video features into a more efficient representation, we further propose a MQ-former to aggregate them into a set of learnable queries. Ours MQ-former is inspired by the Q-Former in BLIP-2 [56] and augments its content queries q_{con} with sentence queries q_{sen} and box queries q_{box}. q_{sen} and q_{box} are obtained by transforming the corresponding prompt features G_{sen} and G_{box} with two separate linear layers. We add q_{sen} and q_{box} to q_{con} to incorporate semantic and positional cues [63]. Note that the use of different types of queries not only enables our method to adapt to a variety of video tasks but also explicitly integrates guidance information from prompts into the visual features.

With this, we begin by splitting the video features F along the temporal dimension, resulting in a sequence of frame features $\{F_i\}_{i=1}^{T^f}$, and then send them to MQ-former in parallel. Within the MQ-former, the summed queries interact with one another, and F_i, through self-attention and cross-attention in an iterative manner, which integrates the frame features into the compact queries. Subsequently, we feed the per-frame queries to a transformer layer [28] for temporal modeling, yielding $Q \in \mathcal{R}^{T^f N_q \times C_q}$, where N_q is the number of queries and set to 32 following the configuration in BLIP-2 [56], C_q represents the feature dimension.

Visual Translator. Alignment between video and textual representations is extremely important to ensure that the output of our model is intrinsically relevant to the video content. To accomplish this, we input Q to a Multi-Layer Perceptron (MLP) layer to project it to the textual embedding space, thereby aligning its dimension with the prompt features G. After this, they are concatenated along the sequence dimension to obtain the multimodal tokens $M \in \mathcal{R}^{(L^g+T^f N_q)\times C^g}$.

Video-grounded Token Decoding. Finally, we employ a token decoder to predict a sequence of tokens based on M. The architecture of our token decoder is similar to popular language decoders [53, 87], with causal self-attention for autoregressive token generation.

3.3. Unified Training and Inference

Training. Conditioned on M, OmniViD is trained to maximize the log-likelihood between the predicted tokens $\hat{y}$ and the target tokens y with cross-entropy loss:

$$\text{maximize} \sum_{k=1}^{L} \log \mathrm{P}(\hat{\mathrm{y}}_{\mathrm{k}}|\mathrm{M}, \mathrm{y}_{1:\mathrm{k}-1}), \qquad (1)$$

where P denotes the softmax probability and L is the length of y. Note that the output of various video tasks could be represented as a sequence of tokens in the unified vocabulary introduced in Sec. 3.1.

Inference. During inference, we predict each token according to the model likelihood, *i.e.*, $P(y_k|M, y_{1:k-1})$, and employ the beam search strategy [35] since it leads to the better performance than argmax sampling or nucleus sampling [41]. Similar to language models, the end of sequence generation is indicated by an EOS token. The event segments for dense video captioning and bounding boxes for visual object tracking could be easily obtained by dequantizing the *time* or *box tokens*.

4. Experiments

4.1. Implementation Details

Datasets. Our training corpus include action recognition datasets (Kinetics-400 [46] and Something-Something V2 [39]), clip captioning datasets (MSRVTT [107] and MSVD [106]), video question answering datasets (MSRVTT [107] and MSVD [106]), dense video captioning datasets (ActivityNet [10]), and visual object tracking datasets (TrackingNet [69], LaSOT [29], GOT10K [43]).

Model Instantiation. We adopt VideoSwin pretrained on Kinetics-600 [13] as the video encoder, and initialize the language encoder and token decoder with pretrained Bart-base [53] model that owns ∼140M parameters. The number of time and box tokens are set to 300 and 1000, respectively. Following BLIP-2 [56], we adopt the same architecture of Bert-Base for our MQ-Former, which consists of 12 transformer layers with additionally inserted cross-attention blocks. The positional encodings are added to the outputs of MQ-Former to inject temporal information.

Table 2. Comparison with state-of-the-art video action recognition methods. Note that for MoViNet, we report the best results on both datasets, *i.e.*, A6 on K400 and A3 on SSV2.

Method	K400		SSV2	
	# Frames	Top1	# Frames	Top1
I3D [46]	N/A	72.1	-	-
R(2+1)D-TS [85]	N/A	73.9	-	-
SlowFast [33]	8 × 3 × 10	77.9	-	-
ip-CSN [86]	32 × 3 × 10	79.2	-	-
X3D-XL [31]	16 × 3 × 10	79.1	-	-
SlowFast+NL [33]	16 × 3 × 10	79.8	-	-
CorrNet [90]	32 × 3 × 10	81.0	-	-
MoViNet [48]	120 × 1 × 1	81.5	120 × 1 × 1	64.1
ViT-B-VTN [72]	250 × 1 × 1	78.6	-	-
MViT-B [30]	32 × 1 × 5	80.2	64 × 3 × 1	67.7
XViT [9]	16 × 3 × 1	80.2	32 × 3 × 1	65.4
ViViT-L [2]	16 × 3 × 4	80.6	16 × 3 × 4	65.4
TimeSformer-L [3]	96 × 1 × 3	80.7	96 × 3 × 1	62.3
Mformer-HR [74]	16 × 3 × 10	81.1	16 × 3 × 1	67.1
VideoSwin-B [64]	32 × 3 × 4	82.7	32 × 3 × 1	69.6
UniFormer-B [57]	32 × 1 × 4	82.9	32 × 3 × 1	71.2
Ours	32 × 3 × 4	**83.6**	32 × 3 × 1	**71.3**

Training and Inference Procedures. For the clip-based tasks, including action recognition (AR), clip captioning (CC), and video question answering (ViQA), we sample 32 frames randomly during training and uniformly during inference. For dense video captioning (DVP), we follow [113] to extract frames at 1FPS, and subsample or pad the frame sequence to 160 during both training and inference. For visual object tracking (VOT), we randomly sample two frames in a video sequence during training, following the common practice [20, 102].

We train our model for 50, 20, 50, and 500 epochs for AR, CC, ViQA, DVP, and VOT, respectively. Note that we follow [21, 102] to train VOT for a longer time since the scale of tracking datasets is much larger. Different batch sizes are adopted, *i.e.*, 64 for AR, 8 for CC, 256 for ViQA, 8 for DVP, and 16 for VOT. The model is optimized with the AdamW optimizer [65], with an initial learning rate 5e-6 and decayed to 0 with the cosine scheduler. The frame resolution that we adopt is 224 × 224, augmented with random resized cropping and horizontal flipping. During inference, we average the logits of the generated tokens as the final score for AR to support multi-clip&crop evaluation, and VOT for template update [21, 102]. The threshold for VOT template update is 0.03.

Table 3. Comparison with state-of-the-art video captioning methods on MSRVTT and MSVD. Off-the-shelf object detectors are used for the results marked with †.

Method	MSRVTT				MSVD			
	B@4	M	R	C	B@4	M	R	C
PickNet [22]	41.3	27.7	59.8	44.1	52.3	33.3	69.6	76.5
SibNet [62]	40.9	27.5	60.2	47.5	54.2	34.8	71.7	88.2
OA-BTG† [118]	41.4	28.2	-	46.9	56.9	36.2	-	90.6
GRU-EVE† [1]	38.3	28.4	60.7	48.1	47.9	35.0	71.5	78.1
MGSA [15]	42.4	27.6	-	47.5	53.4	35.0	-	86.7
POS+CG [89]	42.0	28.2	61.6	48.7	52.5	34.1	71.3	88.7
POS+VCT [42]	42.3	29.7	62.8	49.1	52.8	36.1	71.8	87.8
SAAT [124]	39.9	27.7	61.2	51.0	46.5	33.5	69.4	81.0
STG-KD† [73]	40.5	28.3	60.9	47.1	52.2	36.9	73.9	93.0
PMI-CAP [17]	42.1	28.7	-	49.4	54.6	36.4	-	95.1
ORG-TRL† [121]	43.6	28.8	62.1	50.9	54.3	36.4	73.9	95.2
OpenBook [123]	42.8	29.3	61.7	52.9	-	-	-	-
SwinBERT [61]	41.9	29.9	62.1	53.8	58.2	41.3	77.5	120.6
Ours	**44.3**	**29.9**	**62.7**	**56.6**	**59.7**	**42.2**	**78.1**	**122.5**

Table 4. Accuracy (%) of ViQA on MSRVTT and MSVD, Pre VLData: pertaining vision-language data.

Method	PreTrain VLData	MSRVTT	MSVD
ClipBERT [52]	5.6M	37.4	-
CoMVT [79]	100M	39.5	42.6
JustAsk [112]	69M	41.5	46.3
ALIPRO [55]	5.5M	42.1	45.9
OmniVL [92]	18M	44.1	51.0
HCRN [50]	-	35.6	36.1
JustAsk [112]	-	39.6	41.2
Ours	-	**42.3**	**47.7**

4.2. Main Results

1) Action Recognition, as one of the most representative video understanding tasks, aims to identify the action categories in a video. We evaluate the Top-1 accuracy of OmniViD on commonly used datasets, including Kinetics-400 (K400) [46] which consists of 306k short video clips of 400 action categories, and Something-Something V2 (SSV2) [39] which comprises 220k videos of 174 categories. The comparison results with other methods are shown in Table 2. OmniViD achieves the best performance on both datasets, *i.e.*, 83.6% on K400 and 71.3% on SSV2, surpassing VideoSwin [64] by 0.9 and 1.7, respectively. This highlights the advantage of our method.

2) Video Captioning expects the model to generate a textual description for a given video, which simultaneously evaluates the visual comprehension and text generation capability of our method. MSRVTT [107] and MSVD [14], two large-scale open domain video captioning datasets, are adopted and the results are shown in Table 3. We can see that OmniViD outperforms existing models by a clear margin (+2.8 and +1.9 in terms of CIDEr on MSRVTT and MSVD), even if several of them, *e.g.*, OA-BTG [118] and ORG-TRL [121], leverage object detector [38, 40] to extract object information in an offline manner.

Table 5. Dense captioning on the ActivityNet Captions validation set. * denotes pretraining on large-scale video-language dataset YT-Temporal-1B [117].

Method	Captioning			Event Loc.		Overall
	B4	M	C	R	P	SODA_c
DCE [49]	0.17	5.69	12.43	-	-	-
DVC [58]	0.73	6.93	12.61	-	-	-
TDA-CG [91]	1.31	5.86	7.99	-	-	-
SDVC [70]	-	6.92	-	55.58	57.57	-
PDVC [99]	1.65	7.50	25.87	55.42	58.07	5.3
UEDVC [119]	-	-	-	**59.00**	60.32	5.5
Vid2seq [113]	-	-	18.80	-	-	5.4
Vid2seq* [113]	-	8.50	30.10	52.70	53.90	5.8
Ours	**1.73**	**7.54**	**26.00**	45.08	**60.43**	**5.6**

Table 6. Comparisons with the visual object tracking models on LaSOT and TrackingNet.

Method	LaSOT			TrackingNet		
	Suc	P_{norm}	P	Suc	P_{norm}	P
SiamFC [4]	33.6	42.0	33.9	57.1	66.3	53.3
ATOM [26]	51.5	57.6	50.5	70.3	77.1	64.8
SiamPRN++ [54]	49.6	56.9	49.1	73.3	80.0	69.4
DiMP [5]	56.9	65.0	56.7	74.0	80.1	68.7
KYS [6]	55.4	63.3	-	74.0	80.0	68.8
Ocean [120]	56.0	65.1	56.6	-	-	-
AutoMatch [122]	58.2	-	59.9	76.0	-	72.6
PrDiMP [27]	59.8	68.8	60.8	75.8	81.6	70.4
TrDiMP [97]	63.9	-	61.4	78.4	83.3	73.1
Siam R-CNN [88]	64.8	72.2	-	81.2	85.4	80.0
TransT [20]	64.9	73.8	69.0	81.4	86.7	80.3
Unicorn [109]	68.5	76.6	74.1	83.0	86.4	82.2
KeepTrack [67]	67.1	77.2	70.2	-	-	-
STARK [108]	67.1	77.0	-	82.0	86.9	-
AiATrack [37]	-	79.4	73.8	-	87.8	80.4
OSTrack [115]	-	78.7	75.2	-	87.8	82.0
MixFormer [25]	69.2	78.7	74.7	83.1	88.1	81.6
SeqTrack [21]	69.9	79.7	76.3	83.3	88.3	82.2
ARTrack [102]	70.4	79.5	76.6	84.2	88.7	83.5
UNINEXT [110]	72.4	**80.7**	**78.9**	**85.1**	88.2	**84.7**
Ours	**70.8**	79.6	**76.9**	83.8	**88.9**	83.2

3) Video Question Answering aims to answer a natural language question based on the video content. We compare the accuracy of OmniViD with other ViQA models on MSRVTT [107] and MSVD [106] in Table 4. The results demonstrate that OmniViD outperforms both QA-specific

methods, *e.g.*, JustAsk [112], and pertaining methods, *e.g.*, ALIPRO [55], showcasing the effectiveness of our method for complex multimodal reasoning.

4) Dense Video Captioning localizes the events in an untrimmed video and generates the corresponding text descriptions for them. Following the practice of previous methods [99, 113], we evaluate OmniViD in three aspects: 1) the average precision (P), average recall (R) across IOU at 0.3, 0.5, 0.7, 0.9 and their harmonic mean for localization. 2) BLEU4 (B4), METEOR (M), and CIDEr (C) for dense captioning. 3) SODA_c for an overall evaluation. The results are reported in Table 5.

Traditional methods, including both two-stage (*e.g.*, DVC [58], SDVC [70]), and one-stage models (*e.g.*, PDVC [99], UEDVC [119]), all employ the pre-extracted features from video backbones [46] without end-to-end training. Compared to them, OmniViD achieves better results on all the metrics, except for Recall. Our underperformance on recall is because traditional methods always apply a fixed number of localization heads to get a large number of false-positive predictions, *e.g.*, 100 for SDVC [70]. Vid2Seq [113] is the first end-to-end framework for dense video captioning. We can see that our method, although slightly inferior to their pre-trained model on YT-Temporal-1B, can significantly outperform them without large-scale pretraining, *i.e.*, 18.80 *vs.* 26.00 in terms of CIDEr. A detailed comparison between OmniViD and Vid2seq can be found in the appendix.

5) Visual Object Tracking estimates the trajectory of a target object given its position in the first frame, which requires a fine-grained understanding of spatial-temporal information. In Table 6, we compare OmniViD with other tracking models on two most representative datasets, LaSOT [29] and TrackingNet [69]. Success (Suc), precision (P), and normalized precision ($\mathrm{P}_{\mathrm{norm}}$) are reported. It is worth mentioning that although SeqTrack [21] and ARTrack [102] also employ the autoregressive framework for object tracking, OmniViD differs from them in twofold aspects. Firstly, we perform tracking on the complete frame, instead of a cropped region. Second, we encode the reference box to the visual feature of the tracking frame through box queries, rather than just using it as a prompt for the token decoder. It can be observed that OmniViD achieves excellent performance on both LaSOT and TrackingNet, *i.e.*, 79.6 and 88.9 in terms of $\mathrm{P}_{\mathrm{norm}}$, which beats most of the previous SOTA methods.

4.3. Ablation Studies

Analysis of Different Components in OmniViD. In Table 7, we conduct ablation experiments to study the effects of the core components in OmniViD: 1) text & box queries in Mixed Qformer: different queries are the core design of our method to adapt to different video tasks and inject reference information into the frame feature. It can be seen from the 1st and 2nd rows that they improve the VQA and VOT performance by 1.9 and 1.4, respectively. 2) temporal encoder: comparing the results in the 3rd and 5th rows, it is evident that the temporal encoder brings remarkable performance gains on all the tasks, validating the temporal modeling is important for video understanding. 3) initializing token decoder with Bart [53]: the results in row 4 demonstrate that the initialization of the token decoder has a greater impact on captioning tasks, stemming from the fact that the training objectives of captioning tasks are inherently more aligned with the pretraining of the token decoder.

Table 7. Ablation studies on different components of OmniViD.

	Model	AR	CC	ViQA	DVP	VOT
1	w/o TextQuery	83.4	56.5	40.4	5.6	79.2
2	w/o BoxQuery	-	-	-	-	78.2
3	w/o TemEnc	82.5	53.3	41.7	5.1	77.6
4	w/o LangInit	81.7	44.4	39.7	4.5	79.0
5	Ours	**83.6**	**56.6**	**42.3**	**5.6**	**79.6**

Open-vocabulary Action Recognition: Compared to the traditional classifier-based methods, OmniViD is more flexible in adapting to the open-vocabulary (OV) setting by appending the category names to the input textual prompt. As shown in Table 8, OmniViD achieves competitive results than existing OV methods without cumbersome designs.

Table 8. Open-vocabulary results on HMDB-51 and UCF101.

Method	Train	HMDB-51	UCF101
ASR [98]	K400	21.8± 0.9	24.4±1.0
ZSECOC [75]	K400	22.6±1.2	15.1±1.7
UR [127]	K400	24.4±1.6	17.5±1.6
E2E [7]	K400	32.7	48.0
Ours	K400	26.3	32.0

Number of Time and Box Tokens. We further try different numbers of time (N_t) and box (N_b) tokens on the localization tasks. As shown in Figure 4, for both types of tokens, increasing the number could first improve the results since the quantization error is reduced accordingly, and finally converges when $N_t \geq 300$ and $N_b \geq 1000$.

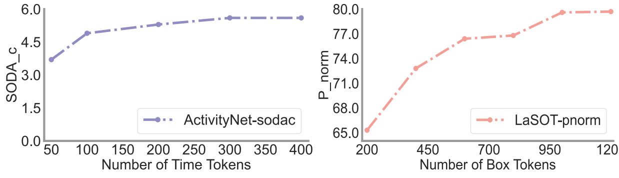


Figure 4. Comparison between joint and separate training.

4.4. Visualizations

We visualize the predictions of OmniViD on various video understanding tasks in Figure 5. From the top two rows, we can see that OmniViD could not only generate accurate and natural captions for videos but also answer questions regarding the characters or activities in the video, showcasing

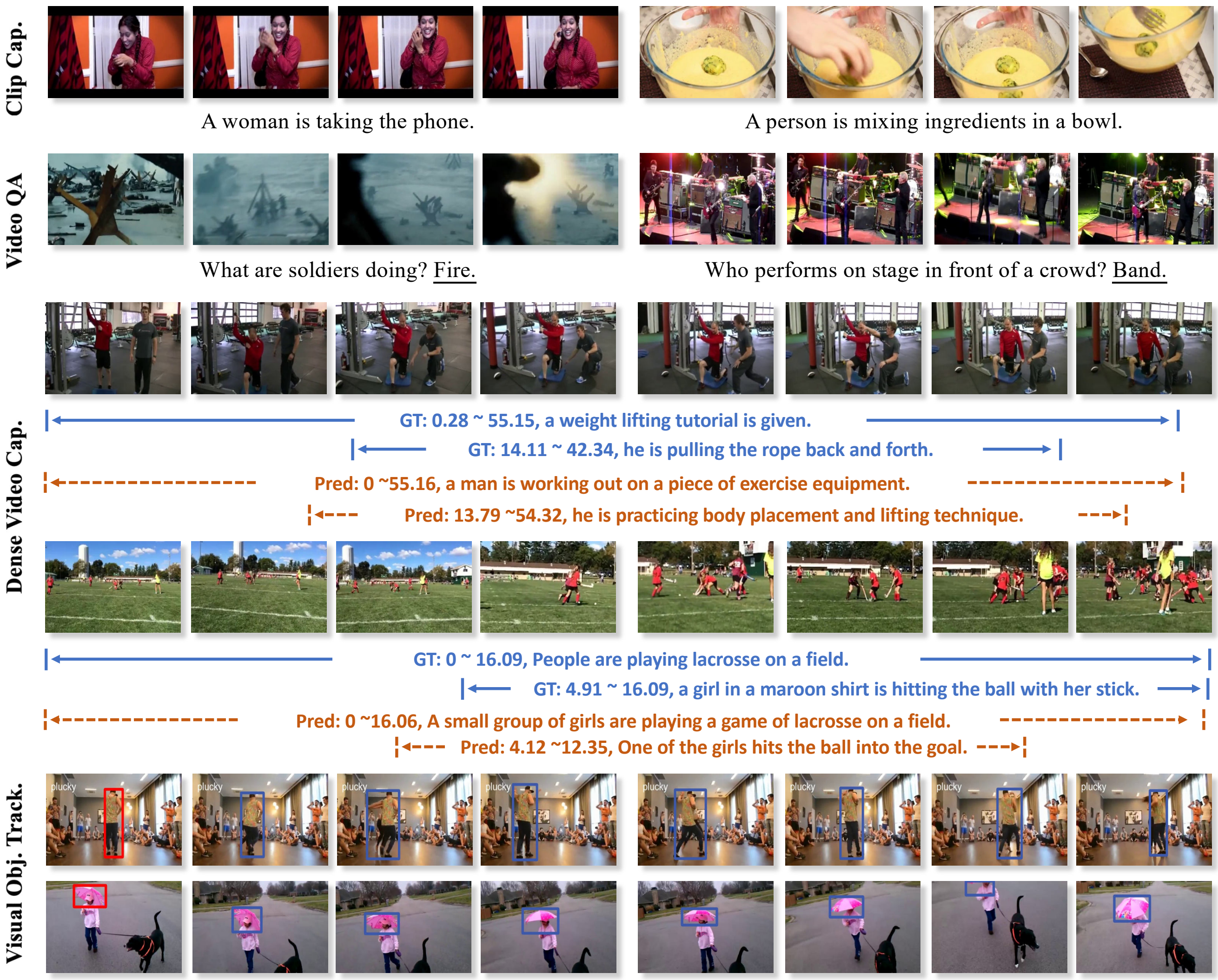

Clip Cap.
A woman is taking the phone.
A person is mixing ingredients in a bowl.
Video QA
What are soldiers doing? Fire.
Who performs on stage in front of a crowd? Band.
Dense Video Cap.
GT: 0.28 ~ 55.15, a weight lifting tutorial is given.
GT: 14.11 ~ 42.34, he is pulling the rope back and forth.
Pred: 0 ~55.16, a man is working out on a piece of exercise equipment.
Pred: 13.79 ~54.32, he is practicing body placement and lifting technique.
GT: 0 ~ 16.09, People are playing lacrosse on a field.
GT: 4.91 ~ 16.09, a girl in a maroon shirt is hitting the ball with her stick.
Pred: 0 ~16.06, A small group of girls are playing a game of lacrosse on a field.
Pred: 4.12 ~12.35, One of the girls hits the ball into the goal.
Visual Obj. Track.
plucky

Figure 5. Visualization of the predictions by OmniViD on different video understanding tasks. From top to down, we show the clip captioning, video question answering, dense video captioning, and visual object tracking visualization results, respectively.

its cross-modal modeling capability. In addition, OmniViD also excels in spatial-temporal localization. The results in 3rd and 4th rows show that it could detect different types of events in videos precisely and produce vivid descriptions for them. Moreover, OmniViD also exhibits remarkable robustness against occlusions and variations in object tracking. These visualizations underscore the versatility and effectiveness of OmniViD across a wide range of video tasks.

5. Conclusion

This paper introduced OmniViD, a generative framework for universal video understanding. We defined a unified output space for different video tasks by supplementing the vocabulary of language models with special *time* and *box tokens*. With this, a wide spectrum of video tasks, including action recognition, clip captioning, video question answering, dense video captioning, and visual object tracking, could be formulated as a video-grounded token generation process, and further, addressed within an encoder-decoder architecture. Extensive experiments on seven prominent video benchmarks showcased the superior video understanding capability and versatility of OmniViD.

Despite the promising results achieved, the joint training performance of OmniViD exhibited some degradation in the spatial-temporal localization tasks compared to separate training. In the future, we will explore more advanced training and optimization strategies on multiple datasets and tasks, to further improve the overall performance and robustness of our method.

Acknowledgement This project was supported by NSFC under Grant No. 62032006 and No. 62102092.

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *CVPR*, 2019. 6

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 5

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1, 2, 5

[4] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCVW*, 2016. 1, 4, 6

[5] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 6

[6] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *ECCV*, 2020. 6

[7] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *CVPR*, 2020. 7

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 3

[9] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. In *NeurIPS*, 2021. 5

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1, 2, 5

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3

[12] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1

[13] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 5

[14] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL-HLT*, 2011. 6

[15] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *AAAI*, 2019. 6

[16] Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. Deep learning for video captioning: A review. In *IJCAI*, 2019. 1

[17] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *ECCV*, 2020. 6

[18] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022. 3

[19] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. In *NeurIPS*, 2022. 3

[20] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 1, 2, 4, 5, 6

[21] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, 2023. 3, 5, 6, 7

[22] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In *ECCV*, 2018. 6

[23] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. In *NeurIPS*, 2021. 1

[24] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *FAI*, 2020. 3

[25] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, 2022. 2, 6

[26] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 6

[27] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 6

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[29] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 5, 7

[30] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *CVPR*, 2021. 1, 2, 5

[31] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 5

[32] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2

[33] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1, 2, 5

[34] Eric D Feigelson, G Jogesh Babu, and Gabriel A Caceres. Autoregressive times series methods for time domain astronomy. *Frontiers in Physics*, 2018. 3

[35] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *ACL*, 2017. 5

[36] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *TMM*, 2017. 1

[37] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *ECCV*, 2022. 6

[38] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2, 6

[39] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 5, 6

[40] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *CVPR*, 2017. 2, 6

[41] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. 5

[42] Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. Joint syntax representation learning and visual cue translation for video captioning. In *ICCV*, 2019. 6

[43] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 2019. 5

[44] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPRW*, 2020. 2

[45] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. *arXiv preprint arXiv:2403.07304*, 2024. 1

[46] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 5, 6, 7

[47] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[48] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *CVPR*, 2021. 5

[49] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1, 6

[50] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, 2020. 2, 6

[51] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2

[52] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 6

[53] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020. 3, 5, 7

[54] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 6

[55] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*, 2022. 6, 7

[56] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4, 5

[57] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. In *ICLR*, 2022. 5

[58] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018. 6, 7

[59] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, 2022. 2

[60] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 1

[61] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, 2022. 1, 2, 6

[62] Sheng Liu, Zhou Ren, and Junsong Yuan. Sibnet: Sibling convolutional encoder for video captioning. In *ACM MM*, 2018. 6

[63] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*, 2022. 4

[64] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 1, 2, 5, 6

[65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[66] Chuofan Ma, Qiushan Guo, Yi Jiang, Ping Luo, Zehuan Yuan, and Xiaojuan Qi. Rethinking resolution in the context of efficient video recognition. In *NeurIPS*, 2022. 2

[67] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *ICCV*, 2021. 6

[68] Christopher Meek, David Maxwell Chickering, and David Heckerman. Autoregressive tree models for time-series analysis. In *ICDM*, 2002. 3

[69] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 1, 2, 5, 7

[70] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, 2019. 6, 7

[71] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 1

[72] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *ICCV*, 2021. 5

[73] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, 2020. 6

[74] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021. 5

[75] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *CVPR*, 2017. 7

[76] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, 2023. 2

[77] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. 1

[78] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. 1

[79] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *CVPR*, 2021. 6

[80] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 1

[81] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *TPAMI*, 2013. 1

[82] Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. Resformer: Scaling vits with multi-resolution training. In *CVPR*, 2023. 1

[83] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[84] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[85] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 5

[86] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 5

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5

[88] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *CVPR*, 2020. 6

[89] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable video captioning with pos sequence guidance based on gated fusion network. In *ICCV*, 2019. 6

[90] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *CVPR*, 2020. 5

[91] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018. 6

[92] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *NeurIPS*, 2022. 1, 2, 6

[93] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*, 2023. 1

[94] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying object tracking by tracking-with-detection. *arXiv preprint arXiv:2303.12079*, 2023. 3

[95] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Chuanxin Tang, Xiyang Dai, Yucheng Zhao, Yujia Xie, Lu Yuan, and Yu-Gang Jiang. Look before you match: Instance understanding matters in video object segmentation. In *CVPR*, 2023. 1

[96] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. Visual tracking with fully convolutional networks. In *ICCV*, 2015. 2, 4

[97] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, 2021. 6

[98] Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *ECML PKDD*, 2017. 7

[99] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, 2021. 2, 6, 7

[100] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 3

[101] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun

Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1, 2

[102] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *CVPR*, 2023. 3, 5, 6, 7

[103] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM MM*, 2015. 1

[104] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data. *TPAMI*, 2024. 1

[105] Zhen Xing, Qi Dai, Zihao Zhang, Hui Zhang, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023. 1

[106] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *CVPR*, 2017. 5, 6

[107] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1, 2, 5, 6

[108] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 2021. 6

[109] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022. 3, 6

[110] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 3, 6

[111] Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *ICCV*, 2023. 2

[112] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 6, 7

[113] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 1, 2, 4, 5, 6, 7

[114] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019. 3

[115] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 6

[116] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *CSUR*, 2006. 1

[117] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022. 6

[118] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *CVPR*, 2019. 6

[119] Qi Zhang, Yuqing Song, and Qin Jin. Unifying event detection and captioning as sequence generation via pre-training. In *ECCV*, 2022. 6, 7

[120] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *ECCV*, 2020. 6

[121] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020. 6

[122] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *ICCV*, 2021. 6

[123] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *CVPR*, 2021. 6

[124] Qi Zheng, Chaoyue Wang, and Dacheng Tao. Syntax-aware action targeting for video captioning. In *CVPR*, 2020. 6

[125] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2

[126] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. 2

[127] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *CVPR*, 2018. 7