Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding

Peng Jin[1,2,3] Ryuichi Takanobu Wancai Zhang[4] Xiaochun Cao[5] Li Yuan[1,2,3*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China
[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China
[4]Nari Technology Co.,Ltd., China [5]School of Cyber Science and Tech., Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
jp21@stu.pku.edu.cn yuanli-ece@pku.edu.cn

Abstract

*Large language models have demonstrated impressive universal capabilities across a wide range of open-ended tasks and have extended their utility to encompass multimodal conversations. However, existing methods encounter challenges in effectively handling both image and video understanding, particularly with limited visual tokens. In this work, we introduce Chat-UniVi, a **Uni**fied **Vi**sion-language model capable of comprehending and engaging in conversations involving images and videos through a unified visual representation. Specifically, we employ a set of dynamic visual tokens to uniformly represent images and videos. This representation framework empowers the model to efficiently utilize a limited number of visual tokens to simultaneously capture the spatial details necessary for images and the comprehensive temporal relationship required for videos. Moreover, we leverage a multi-scale representation, enabling the model to perceive both high-level semantic concepts and low-level visual details. Notably, Chat-UniVi is trained on a mixed dataset containing both images and videos, allowing direct application to tasks involving both mediums without requiring any modifications. Extensive experimental results demonstrate that Chat-UniVi consistently outperforms even existing methods exclusively designed for either images or videos. Code is available at https://github.com/PKU-YuanGroup/Chat-UniVi.*

1. Introduction

Large language models (LLMs), such as GPT-3 [7] and LLaMA [59, 60], showcase substantial universal capabilities that pave the way for achieving general artificial intelligence. However, language represents just one facet of communication. Visual information serves to augment and enhance our comprehension of the world. Therefore, there exists a burgeoning interest in developing a multimodal conversation model that can accommodate various input modalities simultaneously, including images and videos.

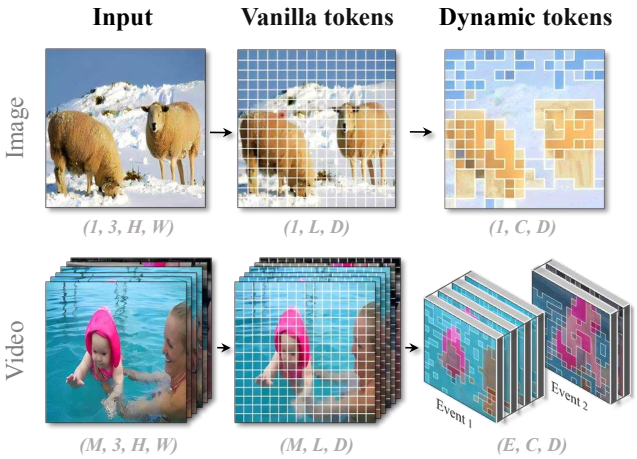


Figure 1. **The unified representation framework for images and videos utilizing a collection of dynamic visual tokens.** "H" and "W" represent the height and width of the input, respectively. "L", "D", "M", "C", and "E" denote the number of vanilla visual tokens, the feature dimension, the frame length, the number of dynamic visual tokens, and the number of events, respectively.

Recent advances in multimodal conversation models, such as MiniGPT-4 [79], LLaVA [38, 39], and mPLUG-Owl [69], focus on integrating visual tokens into LLMs. Despite their commendable progress, existing methods often specialize in either image or video inputs. For instance, methods [38, 39] that prioritize image inputs typically employ a larger number of visual tokens to attain finer spatial understanding. Conversely, methods [44] concentrating on video inputs often compromise spatial comprehension per frame to accommodate more frames for modeling temporal relationships. Although some methods, *e.g.*, Flamingo [1], can extract a fixed number of tokens for each image and

*Corresponding author: Li Yuan.

video using a query transformer, their primary emphasis remains on image understanding, lacking the capability to effectively model temporal comprehension, thus resulting in a limited understanding of videos. Therefore, it is crucial and challenging to enable LLMs for both image and video comprehension within a unified framework.

In this paper, we introduce Chat-UniVi, a **Uni**fied **Vi**sion-language model designed to proficiently comprehend and engage in conversations about both images and videos. Chat-UniVi uniformly represents images and videos using a collection of dynamic visual tokens, enabling it to concurrently capture the spatial details of images and the comprehensive temporal relationship of videos. As illustrated in Fig. 1, images can be depicted through visual tokens of diverse sizes. For example, the primary object, *i.e.*, the sheep in Fig. 1, necessitates a fine-grained representation with numerous visual tokens, while the background, *i.e.*, the snow-capped mountain, can be sufficiently modeled with only one visual token. In the case of videos, the video is initially divided into several events, and subsequently, these visual tokens expand over frames within each event to encapsulate frame-level dynamics. Such unified representation for both images and videos significantly reduces the number of visual tokens while maintaining the expressive capabilities of the model. It is worth noting that longer videos are assigned more visual tokens in our method. Therefore, our method is better suited for variable-length video understanding than existing methods.

To obtain these dynamic visual tokens, we propose a token merging method for progressively merging visual tokens with similar semantic meanings. Specifically, starting with visual tokens initialized by the vision transformer [15], we gradually group them by applying the k-nearest-neighbor based density peaks clustering algorithm, *i.e.*, DPC-KNN [16], on the token features. When it comes to videos, we also utilize DPC-KNN on the frame features to get events. At each merging step, visual tokens assigned to the same cluster are merged by averaging their token features. Finally, we provide a multi-scale representation to the LLMs, where the upper layers of the multi-scale representation encompass high-level semantic concepts, while the lower layers emphasize visual details representations.

The proposed Chat-UniVi has two compelling advantages: **First**, its unified image and video modeling method allows training on the mixed dataset of image and video, enabling direct application to both image and video tasks without any modifications. **Second**, the multi-scale representation contributes to the comprehensive understanding of images and videos, empowering Chat-UniVi to adapt to various tasks, including employing high-level representation for semantic understanding and low-level representation for generating detailed descriptions. We evaluate Chat-UniVi on both image and video understanding tasks. As

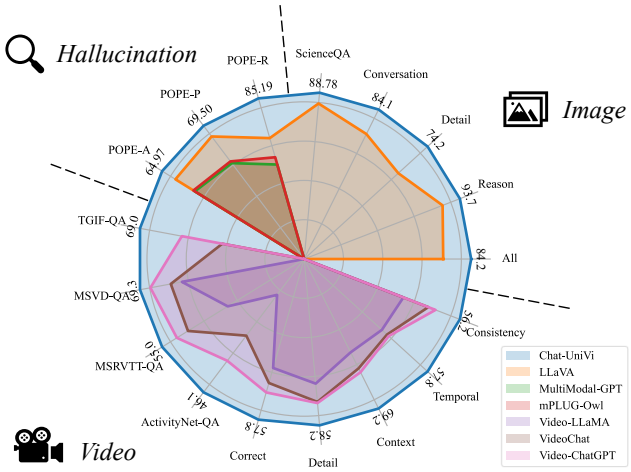


Figure 2. **The proposed Chat-UniVi, designed as a unified model, consistently outperforms even existing methods exclusively designed for either images or videos.** These results demonstrate the advantages of the proposed method.

shown in Fig. 2, compared to other methods focused exclusively on either images or videos, Chat-UniVi consistently demonstrates superiority in comprehending images and videos. Moreover, we also provide evidence of the advantages of joint training of images and videos for multimodal large language models. The main contributions are summarized as follows:

- We propose a unified visual representation for LLMs, enabling LLMs to comprehend both images and videos.
- We uniformly represent images and videos using multi-scale dynamic visual tokens and propose a token merging method to obtain these dynamic visual tokens.
- Without fine-tuning, Chat-UniVi attains competitive performance in both image and video tasks and achieves impressive results in the object hallucination benchmark.

2. Related Work

Large Language Models. Large language models [48, 50, 61] have made disruptive progress, primarily attributed to the expansion of training data and the substantial increase in model parameters. Inspired by the success of GPT-3 [7], numerous LLMs have subsequently been developed, including PaLM [13], OPT [74], BLOOM [54], InstructGPT [47], and ChatGPT [45]. However, language represents just one facet of communication. Visual information serves to augment and enhance our comprehension of the world [5, 23–26, 28, 62, 77]. In this work, we introduce Chat-UniVi, designed to comprehend both image and video inputs.

Large-scale Multimodal Models. Existing large-scale multimodal models [4, 9–11, 17–19, 30, 35–37, 42, 53, 64, 76, 78] can be broadly categorized into two classes. The first class of methods [56, 57, 63, 68] involves using LLMs as a dispatch scheduler, facilitating connections between

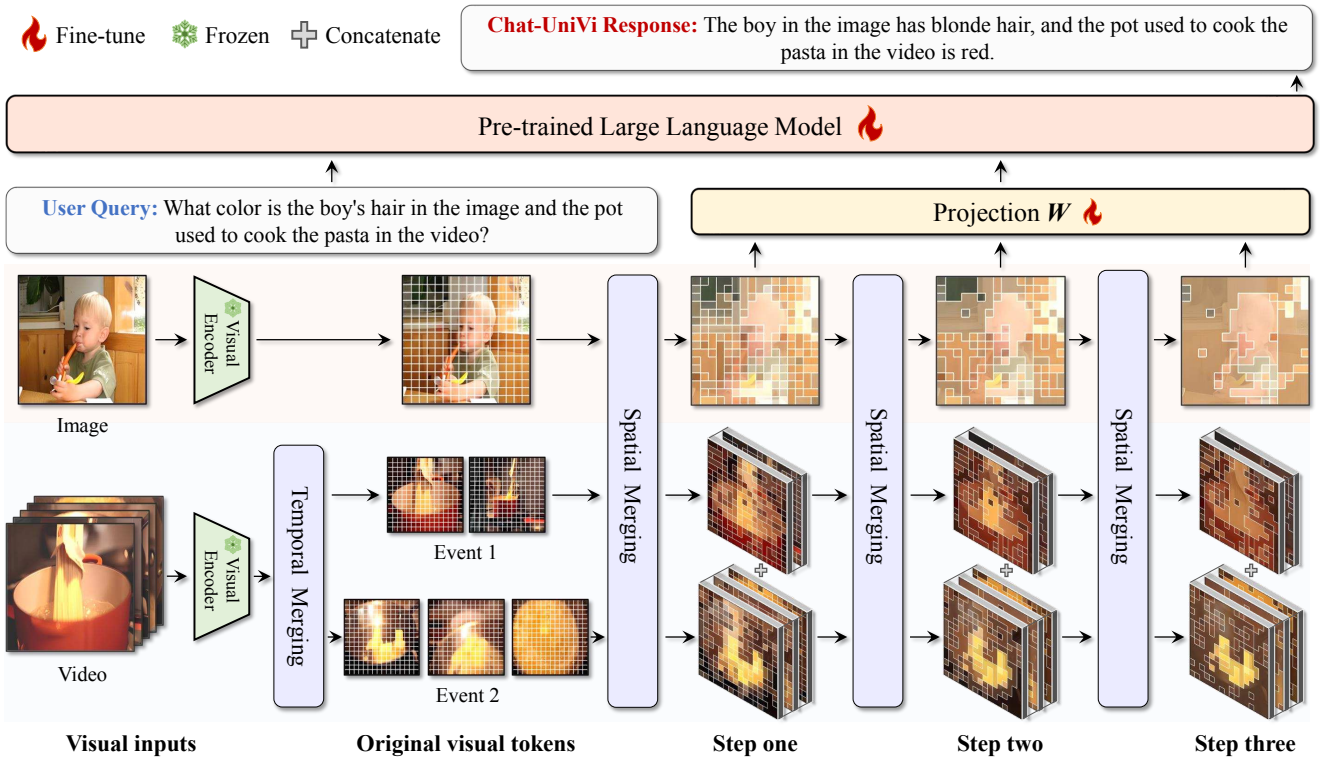


Figure 3. **The overview of the proposed Chat-UniVi for conversations containing both images and videos.** Chat-UniVi uniformly represents images and videos using a collection of dynamic visual tokens and provides a multi-scale representation that equips large language models to perceive both high-level semantic concepts and low-level visual details.

various expert models to handle different vision tasks. The second class of methods [31, 31, 46] emphasizes the integration of models from different modalities into end-to-end trainable models. More recently, there have also been several dedicated multimodal models tailored for video processing, such as Video-LLaVA [34], Video-ChatGPT [44], VideoChat [32], and Video-LLaMA [72]. Despite their commendable progress, existing methods often focus exclusively on either image or video inputs. In this work, we focus on developing an end-to-end trained multimodal model for both image and video tasks. Although Flamingo also supports both image and video inputs, it can only extract a fixed number of tokens for videos of varying lengths with a query transformer. Recent works [9, 64] have explored the use of separately pre-trained image and video encoders for processing, but these methods introduce model redundancy and prove challenging to train together. Hence, it does not align with our focus on achieving a unified vision-language model. In contrast to the previous works, the proposed method uniformly represents images and videos using multi-scale dynamic visual tokens.

Dynamic Visual Token. There have also been recent methods [6, 43, 51, 52, 66, 71] to explore the role of dynamic tokens within the transformer framework. However, none of these methods can be directly extended to video. We summarize the advantages of our method as follows: (i) **Supporting video input.** In contrast to other methods, Chat-UniVi extends the dynamic token method to incorporate video inputs, achieving the integration of image and video representations for the first time. Our work is the first to demonstrate that this unified representation can reconcile the intricate spatial details of images with the broader temporal understanding required for videos. (ii) **Without parameters.** Our clustering method is parameter-free. Interestingly, we find that this parameter-free clustering method serves as the linchpin to the success of our model. We attribute this phenomenon to the gradient instability in multimodal conversation training, which hinders the convergence of parameterized methods. Comparisons of Chat-UniVi and other dynamic token methods are provided in the appendix.

3. Methodology

Chat-UniVi aims to model images and videos concurrently within a language sequence that can be comprehended by Large Language Models (LLMs) in a unified framework. Chat-UniVi achieves this by uniformly representing images and videos through a set of dynamic visual tokens, bridging the intricate spatial details of images with the broader tem-

poral comprehension needed for videos. The overview of the proposed Chat-UniVi is shown in Fig. 3.

3.1. Dynamic Visual Tokens for Image and Video

Building upon the vanilla Vision Transformer, most methods generate equally important visual tokens by dividing the image into regular and fixed grids. However, it is evident that not all regions hold equal significance in vision-language tasks. For example, capturing the background may require only a single visual token. Drawing inspiration from this insight, We amalgamate non-essential tokens to derive dynamic vision regions as input for LLMs.

Spatial Visual Token Merging. For an input image, we adopt the vision encoder of CLIP [49] to provide the original visual tokens $\boldsymbol{Z} = \{z_i\}_{i=1}^{L}$, where L is the number of visual tokens each image is divided into. To amalgamate non-essential visual tokens, we utilize DPC-KNN [16], a k-nearest-neighbor based density peaks clustering algorithm, to cluster the visual tokens. Starting with visual tokens $\boldsymbol{Z} = \{z_i\}_{i=1}^{L}$ initialized by the vision transformer, we first compute the local density ρ_i of each token z_i according to its K-nearest neighbors, which is formulated as:

$$\rho_i = \exp\Big(-\frac{1}{K}\sum_{z_k \in \mathrm{KNN}(z_i, \boldsymbol{Z})} \|z_k - z_i\|^2\Big), \tag{1}$$

where $\mathrm{KNN}(z_i, \boldsymbol{Z})$ is the K-nearest neighbors of z_i in $\boldsymbol{Z}\backslash\{z_i\}$. "$Z\backslash\{z_i\}$" denotes removing $\{z_i\}$ from $\boldsymbol{Z}$. Intuitively, ρ_i denotes the local density of token z_i. Then, we compute the distance index δ_i of the token z_i:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \|z_j - z_i\|^2, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i. \\ \max\limits_{j} \|z_j - z_i\|^2, & \text{otherwise.} \end{cases} \tag{2}$$

In essence, δ_i represents the distance between the given token z_i from other high-density tokens. We identify those tokens with relatively high $\rho_i \times \delta_i$ as cluster centers and then allocate other tokens to their nearest cluster center according to the Euclidean distances. Finally, we utilize the average token within each cluster to represent the corresponding cluster. The vision region of the merged token is the union of the vision regions within the corresponding cluster.

Temporal Visual Token Merging. To adapt the dynamic visual tokens to video inputs, we extend the visual tokens across frames. However, directly consolidating all frames into a limited number of visual tokens may lead to the loss of temporal information within the video. For example, in Fig. 3, the video demonstrates the process of cooking pasta before preparing the sauce. Simply merging all frames would pose challenges for the model in determining the correct sequence, such as whether to prepare the sauce first, cook the pasta first, or simultaneously cook the pasta while preparing the sauce. Therefore, we propose temporal visual token merging to first divide the video into several critical events. Subsequently, we make the visual tokens only expand over frames within the same event.

Given the m_{th} frame $\boldsymbol{Z}^m = \{z_i^m\}_{i=1}^{L}$ of a video, we first apply mean-pooling over all tokens to obtain the frame-level representation f^m. Similar to the spatial visual token merging method, we leverage DPC-KNN to amalgamate non-essential frames. Specifically, we first compute the local density ρ^m and the distance index δ^m of each frame f^m. Frames with relatively high $\rho^m \times \delta^m$ are identified as cluster centers, and other frames are then assigned to their nearest cluster center based on Euclidean distances. We treat each cluster as a critical event and denote the set of indexes of the frames in the cluster as $\boldsymbol{F}$. Therefore, the set of visual tokens within the n_{th} event $\boldsymbol{F}_n$ can be formulated as:

$$\tilde{\boldsymbol{Z}}_n = \big\{z_i^m | m \in \boldsymbol{F}_n,\ i \in \{1, 2, ..., L\}\big\}. \tag{3}$$

After completing the temporal visual token merging, we obtain the set of visual tokens within the event, *i.e.*, $\tilde{\boldsymbol{Z}}$. To make the visual tokens expand over frames within the event, we adjust Eq. (1) and Eq. (2) in the spatial visual token merging method to the following form:

$$\begin{aligned} \tilde{\rho}_i &= \exp\Big(-\frac{1}{K}\sum_{z_k \in \mathrm{KNN}(z_i, \tilde{\boldsymbol{Z}})} \|z_k - z_i\|^2\Big), \\ \tilde{\delta}_i &= \begin{cases} \min\limits_{j:\tilde{\rho}_j > \tilde{\rho}_i} \|z_j - z_i\|^2, & \text{if } \exists j \text{ s.t. } \tilde{\rho}_j > \tilde{\rho}_i. \\ \max\limits_{j} \|z_j - z_i\|^2, & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

Finally, we concatenate the expanded dynamic visual tokens together in order of events to ensure the broader temporal understanding required for videos.

Multi-scale Representation. To further enhance the capabilities of our model, we propose a multi-step aggregation method designed to provide multi-scale visual features for LLMs. Specifically, in Chat-UniVi, the initial visual tokens at the first merging step are derived from the vision encoder of CLIP. Then, we progressively merge visual tokens with similar semantic meanings and obtain different numbers of tokens in different steps. The higher-level features encompass abstract semantic concepts, while the lower levels emphasize representations of visual details. In practice, we execute a three-step aggregation process for each input image or video. Finally, we concatenate the outputs from each merging step and utilize a trainable projection matrix $\boldsymbol{W}$ to transform these multi-scale visual features into language embedding tokens, which serve as inputs for LLMs.

It is worth noting that despite the concatenation, the number of visual tokens in our method remains significantly lower than the original visual tokens generated by the vision transformer. For example, while LLaVA [39] uses 256 visual tokens, our method utilizes only 112 visual tokens.

Methods	LLM Size	Visual Tokens	Conversation	Detail	Reason	All
LLaVA [39]	13B	256	83.1	75.3	96.5	85.1
LLaVA [39]	7B	256	70.3	56.6	83.3	70.1
LLaVA [39]†	7B	256	78.8	70.2	91.8	80.4
Chat-UniVi	7B	**112**	**84.1**	**74.2**	**93.7**	**84.2**

Table 1. **GPT-based evaluation for image understanding.** "†" denotes our own re-implementation of LLaVA under our training settings (same foundation model, same image data, and same training scheme) for a fair comparison.

Methods	LLM Size	Correct	Detail	Context	Temporal	Consistency
Video-LLaMA [72]	7B	39.2	43.6	43.2	36.4	35.8
LLaMA-Adapter [73]	7B	40.6	46.4	46.0	39.6	43.0
VideoChat [32]	7B	44.6	50.0	50.6	38.8	44.8
Video-ChatGPT [44]	7B	48.0	50.4	52.4	39.6	47.4
Chat-UniVi	7B	**57.8**	**58.2**	**69.2**	**57.8**	**56.2**

Table 2. **GPT-based evaluation for video understanding.** The results reported in Maaz et al. [44] span a range from 0 to 5. To standardize the metrics, we normalize all scores to a scale of 0 to 100.

Methods	LLM Size	Subject			Context Modality			Grade		Average
		NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Random Choice [41]	-	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Human [41]	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
Zero-shot Question Answering Accuracy (%)										
GPT-4 [39]	1T+	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
GPT-3 [41]	175B	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87	74.04
LLaVA [39]†	7B	47.78	41.96	53.64	47.90	44.03	51.92	49.63	45.29	48.08
Chat-UniVi	7B	**58.61**	**61.08**	**61.82**	**57.33**	**58.25**	**61.39**	**62.04**	**56.23**	**59.96**
Fine-tuning Question Answering Accuracy (%)										
LLaVA [39]	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA [39]†	7B	79.71	91.68	82.82	80.94	83.24	81.46	83.74	81.74	83.02
LLaMA-Adapter [73]	7B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaMA-SciTune [20]	7B	84.50	**94.15**	82.91	88.35	83.64	**88.74**	85.05	85.60	86.11
Chat-UniVi	7B	**88.50**	93.03	**85.91**	**88.51**	**85.97**	88.15	**88.88**	**88.60**	**88.78**

Table 3. **Zero-shot and fine-tuning question answering accuracy on the ScienceQA test set.** Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. "†" denotes our own re-implementation of LLaVA under our training settings for a fair comparison.

3.2. Multimodal Training Scheme

Multimodal Pre-training. Following the approach of previous works [39], our training is divided into two stages. In the first stage, we pre-train the projection matrix $\boldsymbol{W}$ while freezing both the LLM and the vision encoder. This strategic freezing of the LLM empowers our method to effectively capture semantic visual information without any discernible compromise in the performance of LLMs.

Joint Instruction Tuning. After completing the first stage, the model is able to understand human queries but still fails to generate reasonable and coherent linguistic responses. In the second stage, we fully fine-tune the large language model and the projection matrix $\boldsymbol{W}$ on a multimodal instruction-following dataset. This dataset is a composite of multi-turn conversations and single-turn conversations presented in a conversational format, alongside single images, multiple images, and videos as visual input. Through joint training on the mixture dataset, Chat-UniVi achieves a superior comprehension of various directives and produces more natural and dependable output. Moreover, it exhibits the distinctive ability to seamlessly process both images and videos without requiring any realignment.

4. Experiments

4.1. Experimental Setup

Model Settings. We adopt the vision encoder of CLIP (ViT-L/14) [49] as the visual foundation model. Besides, we chose the Vicuna-v1.5 model [58], which consists of 7B parameters, as our language foundation model.

Data and Training Details. For the multimodal pre-training stage, we utilize the image-caption pairs from various datasets, including COCO [12] and CC3M-595K screened from CC3M [55] by LLaVA [39]. We pre-train Chat-UniVi for one epoch with a batch size of 128, employing the AdamW [27, 40] optimizer with a cosine schedule. The learning rate is set to 2e-3, and the warm-up rate is 0.03. For the joint instruction tuning stage, we incorporate multimodal instruction data from multiple sources: (i) multimodal in-context instruction datasets, such as MIMIC-IT [2, 21, 29], (ii) visual instruction datasets, such as LLaVA, (iii) video instruction data from Video-ChatGPT [44]. All input images or frames are resized to 224×224. We train Chat-UniVi for 2 epochs with a batch size of 128, and the learning rate is set to 2e-5.

Methods	LLM Size	MSRVTT-QA		MSVD-QA		TGIF-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [67]	1B	16.8	-	32.2	-	41.0	-	24.7	-
Video-LLaMA [72]	7B	29.6	1.8	51.6	2.5	-	-	12.4	1.1
LLaMA-Adapter [73]	7B	43.8	2.7	54.9	3.1	-	-	34.2	2.7
VideoChat [32]	7B	45.0	2.5	56.3	2.8	34.4	2.3	26.5	2.2
Video-ChatGPT [44]	7B	49.3	2.8	64.9	3.3	51.4	3.0	35.2	2.7
Chat-UniVi	7B	**55.0**	**3.1**	**69.3**	**3.7**	**69.0**	**3.8**	**46.1**	**3.3**

Table 4. **Zero-shot video question answering accuracy.** We follow the evaluation protocol in Maaz et al. [44], *i.e.*, employing GPT-assisted evaluation to assess the capabilities of models. "Score" denotes the confidence score from 0 to 5 assigned by the GPT model.

Methods	LLM Size	Random (POPE-R)			Popular (POPE-P)			Adversarial (POPE-A)		
		Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes
LLaVA [39]	13B	64.12	73.38	83.26	63.90	72.63	81.93	58.91	69.95	86.76
MiniGPT-4 [79]	13B	79.67	80.17	52.53	69.73	73.02	62.20	65.17	70.42	67.77
InstructBLIP [14]	13B	88.57	89.27	56.57	82.77	84.66	62.37	72.10	77.32	73.03
MultiModal-GPT [19]	7B	50.10	66.71	99.90	50.00	66.67	100.00	50.00	66.67	100.00
mPLUG-Owl [69]	7B	53.97	68.39	95.63	50.90	66.94	98.57	50.67	66.82	98.67
LLaVA [39]†	7B	72.16	78.22	76.29	61.37	71.52	85.63	58.67	70.12	88.33
Chat-UniVi w/o multi-scale	7B	73.88	79.30	74.63	56.36	69.01	90.83	55.63	68.67	91.63
Chat-UniVi w/ multi-scale	7B	**85.19**	**86.05**	**54.67**	**69.50**	**74.39**	**69.10**	**64.97**	**71.54**	**73.10**

Table 5. **Zero-shot object hallucination evaluation on the COCO validation set.** We report the results of the polling-based object probing evaluation (POPE). "Yes" represents the proportion of positive answers that the model outputs. "†" denotes our own re-implementation of LLaVA under our training settings (same foundation model, same image data, and same training scheme) for a fair comparison.

4.2. GPT-based evaluation

Image Understanding. To quantitatively measure the image understanding capability, we report the GPT-4 evaluation results in Tab. 1. Following Liu et al. [39], Zhang et al. [75], we employ 90 questions based on 30 COCO validation images, covering various aspects, including conversation, detail description (Detail), and complex reasoning (Reason). We utilize the GPT-4 model to evaluate the outputs of the model in these three aspects, as well as provide an overall score. For a comprehensive description of image understanding metrics, please refer to the appendix. As shown in Tab. 1, Chat-UniVi uses fewer visual tokens while achieving superior performance. Notably, our method, even as a 7B model, can achieve the performance level of a 13B model, demonstrating the effectiveness of our method.

Video Understanding. To quantitatively measure the video understanding capability, we report the GPT evaluation results in Tab. 2. Following Maaz et al. [44], we employ a test set based on the ActivityNet dataset [8] and utilize the GPT-3.5 model to assign a relative score to the outputs of the model in the following five aspects: Correctness of Information (Correct), Detail Orientation (Detail), Contextual Understanding (Context), Temporal Understanding (Temporal), and Consistency. Please refer to the appendix for more details. As shown in Tab. 2, Chat-UniVi, even as a unified model, significantly surpasses recently proposed state-of-the-art methods that exclusively focus on video, which demonstrates the effectiveness of our method.

4.3. Question-Answer Evaluation

ScienceQA Performance. ScienceQA [41] is a multi-modal science question-answering dataset comprising 21k multiple-choice questions. Each example in ScienceQA contains a visual context, a textual context, a question, and multiple options. We report both zero-shot and fine-tuning results in Tab. 3. As shown in Tab. 3, Chat-UniVi shows competitive performance across all metrics. Notably, Chat-UniVi outperforms LLaMA-SciTune [20], a model specifically tailored for science question answering, which fully demonstrates the superiority of our method.

Zero-shot Video-question Answering Performance. In Tab. 4, we show the zero-shot video-question answering performance on several commonly used open-ended question-answer datasets, including MSRVTT-QA [65], MSVD-QA [65], TGIF-QA FrameQA [22], and ActivityNet-QA [70]. Our evaluation protocol follows that of Maaz et al. [44], utilizing GPT-assisted evaluation to assess the capabilities of models. As shown in Tab. 4, Chat-UniVi outperforms the recently proposed state-of-the-art methods, *e.g.*, FrozenBiLM [67], across various datasets.

4.4. Object Hallucination Evaluation

In Tab. 5, we report the results of the polling-based object probing evaluation [33] (POPE). For details of the polling-based object probing evaluation, please refer to the

Methods	Image Understanding				Video Understanding				
	Conversation	Detail	Reason	All	Correct	Detail	Context	Temporal	Consistency
Only Image	84.0	69.3	89.3	81.5	43.4	48.6	56.8	45.4	46.2
Only Video	72.7	55.8	71.5	66.8	57.4	**58.8**	69.0	56.4	56.0
Image + Video	45.5	31.3	76.1	50.9	51.2	55.6	64.8	50.0	50.4
Video + Image	79.0	69.2	88.5	79.1	45.6	49.8	58.2	46.4	47.8
Image & Video	**84.1**	**74.2**	**93.7**	**84.2**	**57.8**	58.2	**69.2**	**57.8**	**56.2**

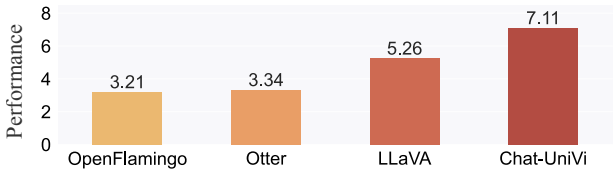
Table 6. **Ablation study about instruction tuning scheme.** "Only Image" indicates training solely on image data. "Image + Video" means training on image data followed by fine-tuning on video data. "Image & Video" denotes training on a mixed dataset.

C_1	C_2	C_3	Visual Tokens	Conversation	Detail	Reason	All
16	8	4	28	78.6	69.0	**95.1**	81.1
32	16	8	56	82.7	67.2	94.5	81.6
64	32	16	112	**84.1**	**74.2**	93.7	**84.2**
128	64	32	224	79.8	68.7	83.8	79.8

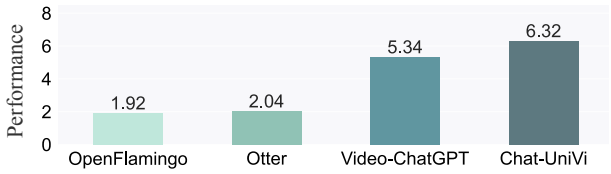
Table 7. **Ablation study about the number of spatial visual clusters.** "C_1", "C_2", and "C_3" denote the number of clusters at the first step, the second step, and the last step, respectively.

Clustering Ratio	Correct	Detail	Context	Temporal	Consistency
$1/M$	51.2	41.8	47.6	32.8	42.2
$1/32$	57.2	58.0	**69.6**	56.2	54.2
$1/16$	**57.8**	**58.2**	69.2	**57.8**	56.2
$1/8$	56.8	**58.2**	68.0	55.8	**57.8**

Table 8. **Ablation study about the number of temporal visual clusters.** "M" is the frame length. "$1/M$" denotes that the model directly consolidates all frames into a single event.



(a) Comparison of image-based conversations



(b) Comparison of video-based conversations

Figure 4. **Human evaluations.** In 30 image conversation scenarios and 30 video conversation scenarios, the evaluators rate the model on a scale of 0 to 10 based on its multimodal conversation performance. Finally, we use the average score as the final model score.

appendix. As shown in Tab. 5, Chat-UniVi outperforms the recently proposed state-of-the-art methods. Moreover, we find that multi-scale representation improves the ability to resist hallucinations. It is worth noting that, as a 7B model, our method even outperforms the 13B model, such as MiniGPT-4. We attribute this success to the multi-scale representation that equips our method to perceive both high-level semantic concepts and low-level visual appearance.

4.5. Ablative Analysis

Effect of the Tuning Scheme. In Tab. 6, we provide the ablation study on the instruction tuning scheme. We find that visual instruction tuning using only one type of medium, such as images, results in a decrease in comprehension of another medium, such as videos. However, pre-training on one medium and fine-tuning on another leads to knowledge degradation from the pre-training stage. In contrast, our joint training strategy, which involves training on a mixed dataset of images and videos, endows the model with the capability to process both types of visual inputs. Among all tuning schemes, joint training consistently achieves the highest performance, confirming its effectiveness.

Effect of the Number of Spatial Visual Clusters. To explore the influence of the number of spatial visual clusters, we provide the ablation results in Tab. 7. We find that a smaller number of visual clusters may decrease the capacity to grasp fine visual details, whereas a larger number of visual clusters may introduce redundancy and potentially reduce the overall performance of the model. To strike a balance between detailed understanding and model learning complexity, we set the number of clusters at the three levels to 64, 32, and 16 respectively in practice.

Effect of the Number of Temporal Visual Clusters. Videos vary in length, with longer videos typically containing more events. Therefore, in Chat-UniVi, the number of temporal visual clusters is determined proportionally based on the number of input video frames. As shown in Tab. 8, we find that a smaller clustering ratio may result in the loss of crucial temporal information within the video. Conversely, a larger clustering ratio increases the computational overhead of the model. We observe that the model performs optimally when the clustering ratio is set to $1/16$. Therefore, in practice, we adopt a default temporal clustering ratio of $1/16$ for optimal performance.

4.6. Qualitative Analysis

Human Evaluation. In our evaluation, we manually assess the performance of Chat-UniVi and baselines in 30

Figure 5. **Visualization of the dynamic visual tokens.** For clarity in observation, we map the dynamic visual tokens of the video back to each frame for visualization. Please refer to the appendix for additional visualizations and conversation examples of our model.

image conversation scenarios and 30 video conversation scenarios. The results are presented in Fig. 4. OpenFlamingo [3], derived from Flamingo [1], and Otter [29], an in-context instruction tuning variant of OpenFlamingo, are also included in our comparison. As shown in Fig. 4, we find that methods based on Flamingo exhibit limitations in their ability to comprehend videos. This limitation is attributed to their use of a query transformer to extract a fixed number of visual tokens from videos of varying lengths, which hinders their effectiveness in modeling temporal comprehension. In contrast, Chat-UniVi, functioning as a unified model, not only outperforms methods built upon the Flamingo but also surpasses models specifically designed for image and video.

Visualization of the Dynamic Visual Tokens. We provide the visualization in Fig. 5 and invite readers to explore more visualizations in the appendix. It is important to emphasize that our proposed token merging method operates without the need for object outline labels. As shown in Fig. 5, the proposed dynamic visual tokens effectively generalize objects and backgrounds. This capability enables Chat-UniVi to reconcile the intricate spatial nuances of images with the broader temporal understanding required for videos with a limited number of visual tokens.

5. Conclusion

In this paper, we introduce Chat-UniVi, a unified multimodal large language model designed to comprehend and engage in conversations about both images and videos. To seamlessly bridge the intricate spatial nuances of images with the broader temporal understanding required for videos, we propose a unified representation framework employing dynamic visual tokens. This representation leverages DPC-KNN to progressively cluster visual tokens and provides multi-scale features. More encouragingly, Chat-UniVi is trained on a mixed dataset encompassing both images and videos, enabling it to be directly applicable to tasks involving both media types without requiring any modifications. Extensive experimental results demonstrate that Chat-UniVi, as a unified model, consistently surpasses even methods exclusively designed for images or videos.

Acknowledgements. This work was supported by the National Key R&D Program of China (2022ZD0118101), Nature Science Foundation of China (No.62202014), and Shenzhen Basic Research Program (No.JCYJ20220813151736001).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736, 2022. 1, 8

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 5

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 8

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 2

[6] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3

[7] 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*, pages 1877–1901, 2020. 1, 2

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 6

[9] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023. 2, 3

[10] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[11] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 6

[15] 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. 2

[16] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99: 135–145, 2016. 2, 4

[17] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. LLaMA-Adapter V2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[18] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. SPHINX-X: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

[19] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. MultiModal-GPT: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 2, 6

[20] Sameera Horawalavithana, Sai Munikoti, Ian Stewart, and Henry Kvinge. Scitune: Aligning large language models with scientific multimodal instructions. *arXiv preprint arXiv:2307.01139*, 2023. 5, 6

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 5

[22] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 6

[23] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *NeurIPS*, pages 30291–30306, 2022. 2

[24] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, pages 2472–2482, 2023.

[25] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video

retrieval with disentangled conceptualization and set-to-set alignment. In *IJCAI*, pages 938–946, 2023.

[26] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *ICCV*, pages 2470–2481, 2023. 2

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Nathan Labiosa, Dat Huynh, and Ser-Nam Lim. Visual information and large language models: A deeper analysis. 2

[29] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 5, 8

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 2

[31] 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. 3

[32] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3, 5, 6

[33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6

[34] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3

[35] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. MoE-LLaVA: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 2

[36] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023.

[37] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 4, 5, 6

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

[41] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022. 5, 6

[42] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. In *NeurIPS*, 2023. 2

[43] Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points. In *ICLR*, 2023. 3

[44] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 3, 5, 6

[45] OpenAI. Introducing chatgpt. *CoRR*, 2022. 2

[46] OpenAI. GPT-4 technical report. *CoRR*, 2023. 3

[47] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, pages 27730–27744, 2022. 2

[48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 4, 5

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[51] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, pages 13937–13949, 2021. 3

[52] Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. TESTA: Temporal-spatial token aggregation for long-form video-language understanding. *arXiv preprint arXiv:2310.19060*, 2023. 3

[53] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023. 2

[54] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 2

[55] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 5

[56] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 2

[57] Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 2

[58] Vicuna Team. Vicuna: An open chatbot impressing gpt-4 with 90% chatgpt quality. 2023. 5

[59] 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

[60] 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

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[62] 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*, pages 5696–5710, 2022. 2

[63] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2

[64] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 2, 3

[65] 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 *ACM MM*, pages 1645–1653, 2017. 6

[66] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18134–18144, 2022. 3

[67] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, pages 124–141, 2022. 6

[68] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2

[69] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 6

[70] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 6

[71] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, pages 11101–11111, 2022. 3

[72] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3, 5, 6

[73] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 5, 6

[74] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

[75] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 6

[76] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. MiniGPT-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 2

[77] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 2

[78] Bin Zhu, Peng Jin, Munan Ning, Bin Lin, Jinfa Huang, Qi Song, Mingjun Pan, and Li Yuan. LLMBind: A unified modality-task integration framework. *arXiv preprint arXiv:2402.14891*, 2024. 2

[79] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 6