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

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

Contents

Abstract This appendix provides additional discussions (Appendix A), implementation details (Appendix B), several additional experiments (Appendix C), additional qualitative analysis (Appendix D), and details of quantitative evaluations (Appendix E).

A. Additional Discussions

A.1. Comparison of Chat-UniVi and Other Multi-modal Methods

Existing methods [8, 14, 15, 22, 32] often focus exclusively on either image or video inputs. Recently, there have also been some methods [1, 6, 27] that support both images and videos, and they can be broadly divided into two classes.

- **Q-former based methods.** The first class of methods uses a query transformer to extract a fixed number of tokens for each image and video. These methods are exemplified by Flamingo [1], OpenFlamingo [3], and Otter [12]. However, videos vary in length, posing a challenge for these methods, as they extract a fixed number of visual tokens from each video, limiting their ability to effectively capture temporal comprehension. Human evaluation results also substantiate that these methods struggle to strike a balance between image and video comprehension.
- **Multi-encoder methods.** The second category of methods employs separate pre-trained image and video encoders to process images and videos independently. Prominent examples of this approach include X-LLM [6] and NExT-GPT [27]. However, these methods introduce redundancy within the model and present difficulties when trained jointly. Most importantly, this approach does not leverage the advantages of joint training with both image and video data. Consequently, they do not align with our primary objective of developing a unified

*Corresponding author: Li Yuan.

Type	Methods	Variable Length Features	Unified Visual Encoder	Benefit from Joint Training
Q-former based methods	Flamingo OpenFlamingo, Otter	✘	✔	–
Multi-encoder methods	X-LLM, NExT-GPT	–	✘	✘
Unified methods	Chat-UniVi	✔	✔	✔

Table A. **Comparison with other methods.** "✘" denotes that the model does not have this property. "✔" denotes that the model has this property. "–" indicates a temporary lack of experimental evidence.

Methods	Parameter-free	Video Input	Image Understanding			
			Conversation	Detail	Reason	All
Ma et al. [17]	✘	✘	71.8	60.9	91.6	75.0
Chat-UniVi	✔	✔	**84.1**	**74.2**	**93.7**	**84.2**

Table B. **Comparison of Chat-UniVi and another token clustering method.** "✘" denotes that the model does not have this property. "✔" denotes that the model has this property.

vision-language model.

In contrast to the previous works, Chat-UniVi uniformly represents images and videos using multi-scale dynamic visual tokens. The proposed Chat-UniVi has two compelling advantages:

- **Variable length video features.** In Chat-UniVi, the number of temporal visual clusters is determined proportionally based on the number of input video frames. In contrast to the Q-former based methods, Chat-UniVi allocates a greater number of visual tokens to longer videos. Therefore, our method is better suited for variable-length video understanding.
- **Unified visual encoder.** Chat-UniVi employs a shared visual encoder to consistently process both images and videos. In contrast to multi-encoder methods, our method eliminates the need for introducing redundant parameters and streamlines the training process.
- **Benefit from joint training.** Due to the unified representation framework for both images and videos, Chat-UniVi can be trained on mixed datasets that include both images and videos. This allows for direct application to tasks involving both images and videos. Most importantly, we find that this joint training strategy can simultaneously enhance the model's understanding of both images and videos.

In Tab. A, we show the comparison of Chat-UniVi and other methods. For Q-former based methods, the advantages of joint training are not shown, and even the performance of the model may affect each other when multiple datasets are mixed [1]. However, the potential to benefit from joint training cannot be ruled out. In addition, the multi-encoder method can also select a video encoder that can encode dynamic length features.

A.2. Comparison of Chat-UniVi and Other Clustering Transformer Methods

There have also been recent methods [11, 17, 28, 29] to explore the role of token clustering within the transformer framework. However, none of these methods can be directly extended to video, and additional parameters need to be trained. We summarize the advantages of our method as follows:

- **Supporting video input.** In contrast to other methods, Chat-UniVi extends the tokens clustering 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.
- **Without parameters.** Our clustering method is parameter-free and therefore requires no training. Interestingly, we find that this parameter-free clustering method serves as the linchpin to the success of our model. As shown in Tab. B, the performance of the clustering method with training parameters is significantly inferior to the parameter-free clustering method we propose. We attribute this phenomenon to the gradient instability in multimodal conversation training, which hinders the convergence of parameterized methods.

A.3. Runtime and Memory Complexity

As shown in Tab. C, the time and memory costs of our clustering algorithm are negligible compared to those of the large language model.

Methods	Time Complexity		Image Inference			Video Inference		
	Spatial	Temporal	Merging (s)	All (s)	Memory (M)	Merging (s)	All (s)	Memory (M)
LLaVA	-	-	0	2.3116	15673	✘	✘	✘
Ours	$\mathcal{O}(L^2D)$	$\mathcal{O}(M^2D)$	0.0027	2.2722	15443	0.0174	4.4040	16533

Table C. **Runtime and memory complexity analysis.** L, D, and M denote the number of vanilla visual tokens, the feature dimension, the frame length, respectively. "✘" denotes that the method does not have this property.

Datasets	Image Inputs	Video Inputs	Multi-turn Conversations	Number of Conversations
Multimodal Pre-training Stage				
CC3M-595K	✔	✘	✘	595K
COCO	✔	✘	✘	956K
Joint Instruction Tuning Stage				
LLaVA-instruct-150K	✔	✘	✔	150K
MIMIC-IT-399K‡	✔	✘	✘	399K
Video-ChatGPT-instruct	✘	✔	✘	100K

Table D. **Description of training data.** "✘" denotes that the dataset does not have this property. "✔" denotes that the dataset has this property. "‡" represents the dataset filtered from MIMIC-IT, containing exclusively image data. In order to further filter the training data, we also delete the duplicate data in LLaVA-instruct-150K and MIMIC-IT.

A.4. Limitations and Future Work

In this section, we delineate the limitations of our work and outline avenues for future research.

The Enduring Impact of Large Language Models. Our method leverages the strength of pre-trained Large Language Models, and as a consequence, also inherits their vulnerabilities.

- **Hallucination.** While our experiments demonstrate the effectiveness of our method in addressing hallucinations, it is important to acknowledge that the issue of hallucinations in LLMs remains a challenge yet to be fully resolved. The phenomenon of illusory responses in LLMs can result in unsupported conjectures during open multimodal conversations, and addressing this issue has the potential to significantly expedite advancements in the field. For a more in-depth exploration of common weaknesses observed in large LLMs, please refer to Brown et al. [4], Rae et al. [21].
- **Long sequence processing.** Transformer-based language models often exhibit suboptimal generalization when confronted with test sequences considerably longer than their training data [19]. This becomes particularly evident in multi-turn conversations, where the model may exhibit forgetfulness of prior conversational context, resulting in erroneous responses. Simultaneously, we find a decline in model performance when multiple videos are inputted, which could also be attributed to constraints associated with sequence length.
- **Prompt sensitivity.** In-context learning has demonstrated disconcerting sensitivity to various aspects of demonstrations, including prompt formats [31]. Notably, different prompt formats can yield entirely contradictory output results. Finding a solution to this issue holds the potential to greatly accelerate progress in the field.

Natural Language Output. Natural language serves as a robust and adaptable input/output interface for describing visual tasks to the model, facilitating the generation of outputs, or estimating conditional probabilities for potential outcomes. However, it may prove to be a less convenient interface for tasks that require conditioning on or predicting more structured outputs, such as bounding boxes, as well as for generating dense pixel predictions. Besides, the flexibility of the natural language output also makes it difficult to evaluate the performance of the model.

More Modalities. Future work can explore alternative modalities, such as audio, in addition to visual inputs. The incorporation of multiple modalities holds the promise of broadening the spectrum of tasks that the model can address, and it has the potential to enhance their performance by leveraging synergies among these various modalities. For example, contemplating audio information alongside video processing can significantly augment the video understanding of the model.

Methods	Image Understanding				Video Understanding				
	Conversation	Detail	Reason	All	Correct	Detail	Context	Temporal	Consistency .
LoRA	76.1	68.6	82.4	75.8	52.8	55.0	63.8	51.6	53.8
Full fine-tuning	**84.1**	**74.2**	**93.7**	**84.2**	**57.8**	**58.2**	**69.2**	**57.8**	**56.2**

Table E. **Comparison between the LoRA and full fine-tuning.** "Detail" denotes the "Detail Description" in the context of image understanding or "Detail Orientation" in the context of video understanding. For image understanding, "Reason" denotes the "Complex Reasoning". For video understanding, "Correct", "Context", and "Temporal" stand for "Correctness of Information", "Contextual Understanding", and "Temporal Understanding", respectively.

Methods	Image Understanding				Video Understanding				
	Conversation	Detail	Reason	All	Correct	Detail	Context	Temporal	Consistency
EVA-CLIP	80.0	74.7	91.2	82.1	57.2	58.8	67.8	55.2	54.6
Openai-CLIP	**84.1**	**74.2**	**93.7**	**84.2**	**57.8**	**58.2**	**69.2**	**57.8**	**56.2**

Table F. **Comparison between the EVA CLIP and the Openai CLIP.** We choose EVA-CLIP (ViT-G), which has a similar number of parameters as Openai-CLIP (ViT-L/14), for the experiment.

Methods	Image Understanding				Video Understanding				
	Conversation	Detail	Reason	All	Correct	Detail	Context	Temporal	Consistency
Single-scale	70.5	63.4	88.3	74.2	54.6	56.4	65.8	52.8	52.2
Multi-scale	**84.1**	**74.2**	**93.7**	**84.2**	**57.8**	**58.2**	**69.2**	**57.8**	**56.2**

Table G. **Ablation study about the multi-scale representation.** These results provide evidence for the benefits of employing a multi-scale representation in multimodal large language models.

B. Implementation Details

B.1. Data Details

For the multimodal pre-training stage, we utilize the image-caption pairs from various datasets, including COCO [7] and CC3M-595K screened from CC3M [23] by LLaVA [16]. All input images are resized to 224×224. 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, 10, 12], (ii) visual instruction datasets, such as LLaVA, (iii) video instruction data from Video-ChatGPT [18]. In order to further filter the training data, we delete the duplicate data in LLaVA-instruct-150K and MIMIC-IT, and delete the video data in MIMIC-IT. 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. For each video, we select 64 frames as input for the model. All input images or frames are resized to 224×224. We provide a detailed description of the training data in Tab. D.

B.2. Model Settings

Following previous works [16], we adopt the vision encoder of CLIP (ViT-L/14) [20] as the visual foundation model. We chose an instruction-tuned variant of LLaMA2 [26], *i.e.*, Vicuna [25], as our language foundation model. Specifically, we utilize the Vicuna-v1.5 model, comprised of 7B parameters.

B.3. Training Hyperparameters

For the multimodal pre-training stage, we pre-train Chat-UniVi for one epoch with a batch size of 128, employing the AdamW 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 train Chat-UniVi for 2 epochs with a batch size of 128, and the learning rate is set to 2e-5, employing the AdamW optimizer with a cosine schedule. The warm-up rate is set to 0.03.

B.4. ScienceQA Fine-tuning Settings

We start with a pre-trained model to fine-tune. We fine-tune the model for 9 epochs with a batch size of 32, employing the AdamW optimizer with a cosine schedule. The learning rate is set to 2e-5, and the warm-up rate is 0.03.

POPE	Methods	LLM Size	Accuracy	Precision	Recall	F1-Score	Yes
Random	Single-scale	7B	73.88	67.03	**97.06**	79.30	74.63
	Multi-scale	7B	**85.19**	**83.59**	88.66	**86.05**	**54.67**
Popular	Single-scale	7B	56.36	53.50	**97.20**	69.01	90.83
	Multi-scale	7B	**69.50**	**64.10**	88.60	**74.39**	**69.10**
Adversarial	Single-scale	7B	55.63	53.07	**97.26**	68.67	91.63
	Multi-scale	7B	**64.97**	**60.23**	88.06	**71.54**	**73.10**

Table H. **Effect of the multi-scale representation on object hallucination.** "Yes" represents the proportion of positive answers that the model outputs.

Methods	Multimodal Pre-training	Instruction Tuning	Image Understanding				POPE-R	Video Inputs
	Datasets	Datasets	Conv	Detail	Reason	All		
LLaVA	CC3M-595K	LLaVA-instruct-150K	82.3	70.2	87.9	80.4	66.83	✘
Chat-UniVi	CC3M-595K	LLaVA-instruct-150K	82.9	68.8	89.8	80.7	82.26	✘
LLaVA	CC3M-595K, COCO	LLaVA-instruct-150K	82.7	68.8	88.8	80.8	72.02	✘
Chat-UniVi	CC3M-595K, COCO	LLaVA-instruct-150K	83.3	72.6	89.0	81.5	82.33	✘
LLaVA	CC3M-595K, COCO	LLaVA-instruct-150K, MIMIC-IT-399K	78.8	70.2	91.8	80.4	74.53	✘
Chat-UniVi	CC3M-595K, COCO	LLaVA-instruct-150K, MIMIC-IT-399K	84.0	69.3	89.3	81.5	83.53	✘
Chat-UniVi w/ video data	CC3M-595K, COCO	LLaVA-instruct-150K, MIMIC-IT-399K, Video-ChatGPT-instruct	**84.1**	**74.2**	**93.7**	**84.2**	**85.19**	✔

Table I. **Ablation of structure and training data.** "✘" denotes that the method does not have this property. "✔" denotes that the method has this property.

C. Additional Experiments

C.1. Comparison between the LoRA and Full Fine-tuning

When the number of model parameters is too large, full fine-tuning of retraining all model parameters becomes expensive, so many recent methods freeze most of the model parameters and train the model with LoRA [9]. We provide the results of the comparison between the LoRA and full fine-tuning in Tab. E. We find that LoRA can achieve competitive performance with full fine-tuning while saving more than half the GPU memory required for training. Future work can use LoRA to extend our method on larger LLMs and vision encoders to achieve better performance.

C.2. Analysis of the Vision Encoder

EVA-CLIP [24] is a recently developed multimodal model with performance comparable to Openai-CLIP [20]. We provide the results of the comparison between EVA-CLIP and Openai-CLIP in Tab. F. We find that the performance of EVA-CLIP is comparable to that of Openai-CLIP when the number of parameters is equal. However, EVA-CLIP offers a larger version of the model with a parameter count of 1.8B, so we think it might be better to adopt a larger EVA-CLIP than Openai-CLIP when using larger LLMs.

C.3. Effect of the Multi-scale Representation

To investigate the impact of the multi-scale representation of our method, we provide the ablation results in Tab. G. Multi-scale representation improves both image understanding and video understanding of the model. These results provide evidence for the benefits of employing a multi-scale representation in multimodal large language models.

C.4. Effect of the Multi-scale Representation on Object Hallucination

Chat-UniVi, as a 7B model, even outperforms the 13B model, *e.g.*, MiniGPT-4, in the object hallucination evaluation. 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. In Tab. H, we show the results of ablation experiments on object hallucination evaluation for the multi-scale representation. We find that multi-scale representation improves the ability to resist hallucinations. Therefore, multi-scale representation is beneficial for multimodal LLMs.

C.5. Ablation of Training Data

We provide comparisons of our method with LLaVA under different conditions in Tab. I. Our method achieves better performance than LLaVA, which we explain in the fol-

POPE	Methods	LLM Size	Accuracy	Precision	Recall	F1-Score	Yes
Random	LLaVA	13B	64.12	59.38	95.99	73.38	83.26
	MiniGPT-4	13B	79.67	78.24	82.20	80.17	52.53
	InstructBLIP	13B	88.57	84.09	95.13	89.27	56.57
	MultiModal-GPT	7B	50.10	50.05	**100.00**	66.71	99.90
	mPLUG-Owl	7B	53.97	52.07	99.60	68.39	95.63
	LLaVA†	7B	72.16	78.22	76.29	78.22	76.29
	Chat-UniVi	7B	**85.19**	**83.59**	88.66	**86.05**	**54.67**
Popular	LLaVA	13B	63.90	58.46	95.86	72.63	81.93
	MiniGPT-4	13B	69.73	65.86	81.93	73.02	62.20
	InstructBLIP	13B	82.77	76.27	95.13	84.66	62.37
	MultiModal-GPT	7B	50.00	50.00	**100.00**	66.67	100.00
	mPLUG-Owl	7B	50.90	50.46	99.40	66.94	98.57
	LLaVA†	7B	61.37	56.63	97.00	71.52	85.63
	Chat-UniVi	7B	**69.50**	**64.10**	88.60	**74.39**	**69.10**
Adversarial	LLaVA	13B	58.91	55.11	95.72	69.95	86.76
	MiniGPT-4	13B	65.17	61.19	82.93	70.42	67.77
	InstructBLIP	13B	72.10	65.13	95.13	77.32	73.03
	MultiModal-GPT	7B	50.00	50.00	**100.00**	66.67	100.00
	mPLUG-Owl	7B	50.67	50.34	99.33	66.82	98.67
	LLaVA†	7B	58.67	54.90	97.00	70.12	88.33
	Chat-UniVi	7B	**64.97**	**60.23**	88.06	**71.54**	**73.10**

Table J. **Detailed results on object hallucination evaluation.** "†" denotes our own re-implementation of LLaVA under our training settings (excluding video data) for a fair comparison.

lowing two aspects. **Multi-scale Representation.** In contrast to LLaVA, which focuses on low-level visual features, our method perceives both high-level semantic concepts and low-level visual details by multi-scale representation. Therefore, our method outperforms LLaVA in conversation, reasoning, and hallucinations. **Scalability.** Our framework supports video input, and by fine-tuning with high-quality video instruction data, the visual capabilities of our models have been significantly enhanced, especially in terms of detailed captioning and reasoning.

Besides, we draw the following two conclusions: (1) Instruction tuning data has a greater impact on performance than pre-training data. (2) High-quality instruction tuning data can significantly enhance model performance. Especially after training on high-quality video data, the performance of the model is greatly improved.

C.6. Detailed Results on Object Hallucination Evaluation

In Tab. J, we report the detailed results of the polling-based object probing evaluation [13]. As shown in Tab. J, Chat-UniVi outperforms the recently proposed state-of-the-art methods. Notably, as a 7B model, our method even outperforms the 13B model, *e.g.*, MiniGPT-4, in the object hallucination evaluation. These results demonstrate the effectiveness of our method.

D. Additional Qualitative Analysis

D.1. Visualization for the Image Inputs

To gain a deeper insight into the functionality of our proposed dynamic visual tokens, we present the additional visualization results for the image inputs in Fig. A. In Fig. A, we provide a diverse range of visualizations encompassing various image categories, including portraits, sports, wildlife, art, architecture, and food. It is crucial to underscore that our proposed token merging method operates without the need for object outline labels and is parameter-free. As shown in Fig. A, the proposed dynamic visual tokens effectively generalize objects and backgrounds, empowering Chat-UniVi to capture the spatial nuances of images using a limited number of visual tokens.

D.2. Visualization for the Video Inputs

To gain a more comprehensive understanding of our proposed dynamic visual tokens, we also present additional visualization results for the video inputs in Fig. B. 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. No-

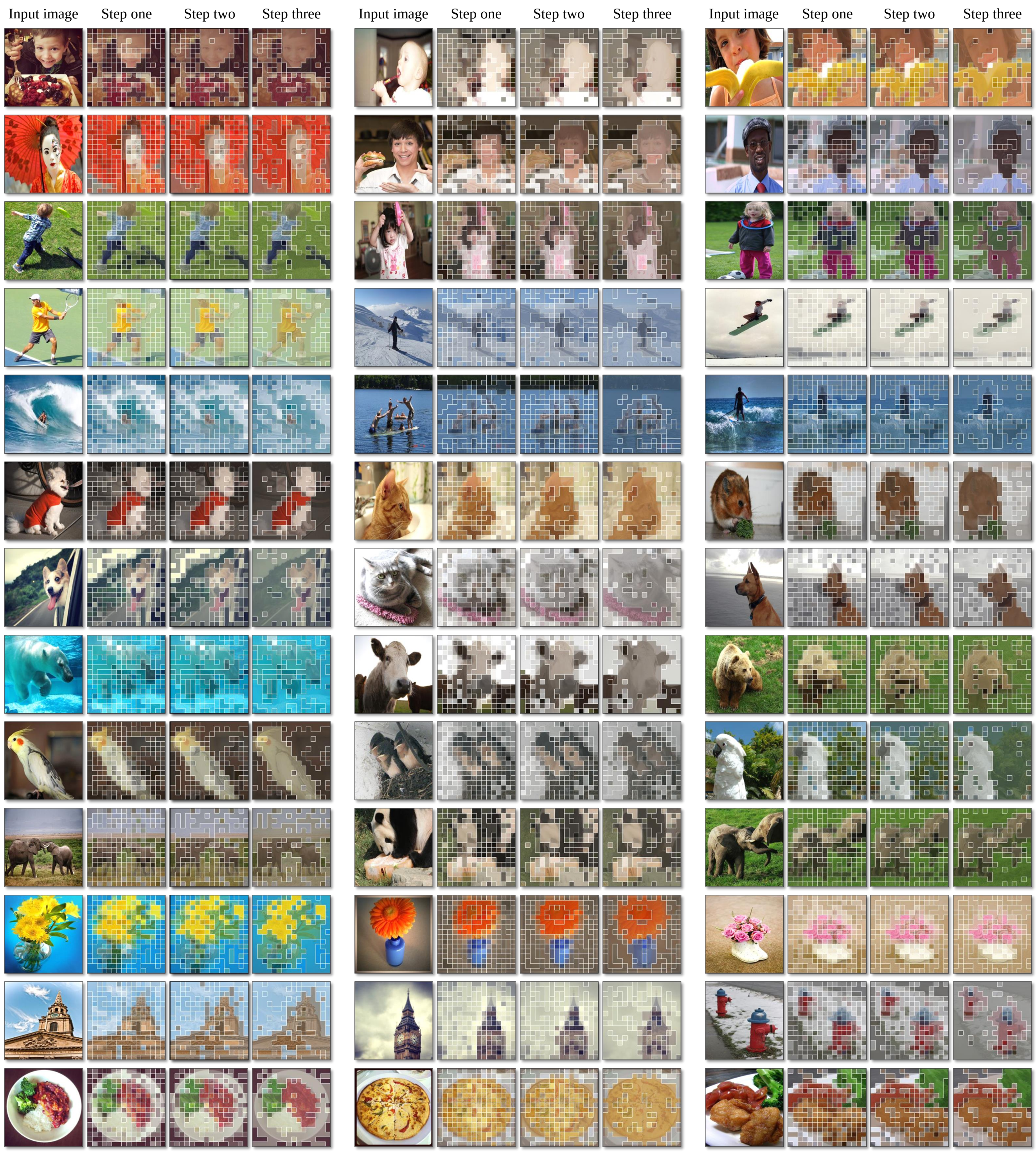


Figure A. **Visualization of the dynamic visual tokens for the image inputs.** We provide a diverse range of visualizations encompassing various image categories, including portraits, sports, wildlife, art, architecture, and food. It is important to emphasize that our proposed token merging method is parameter-free and operates without the need for object outline labels.

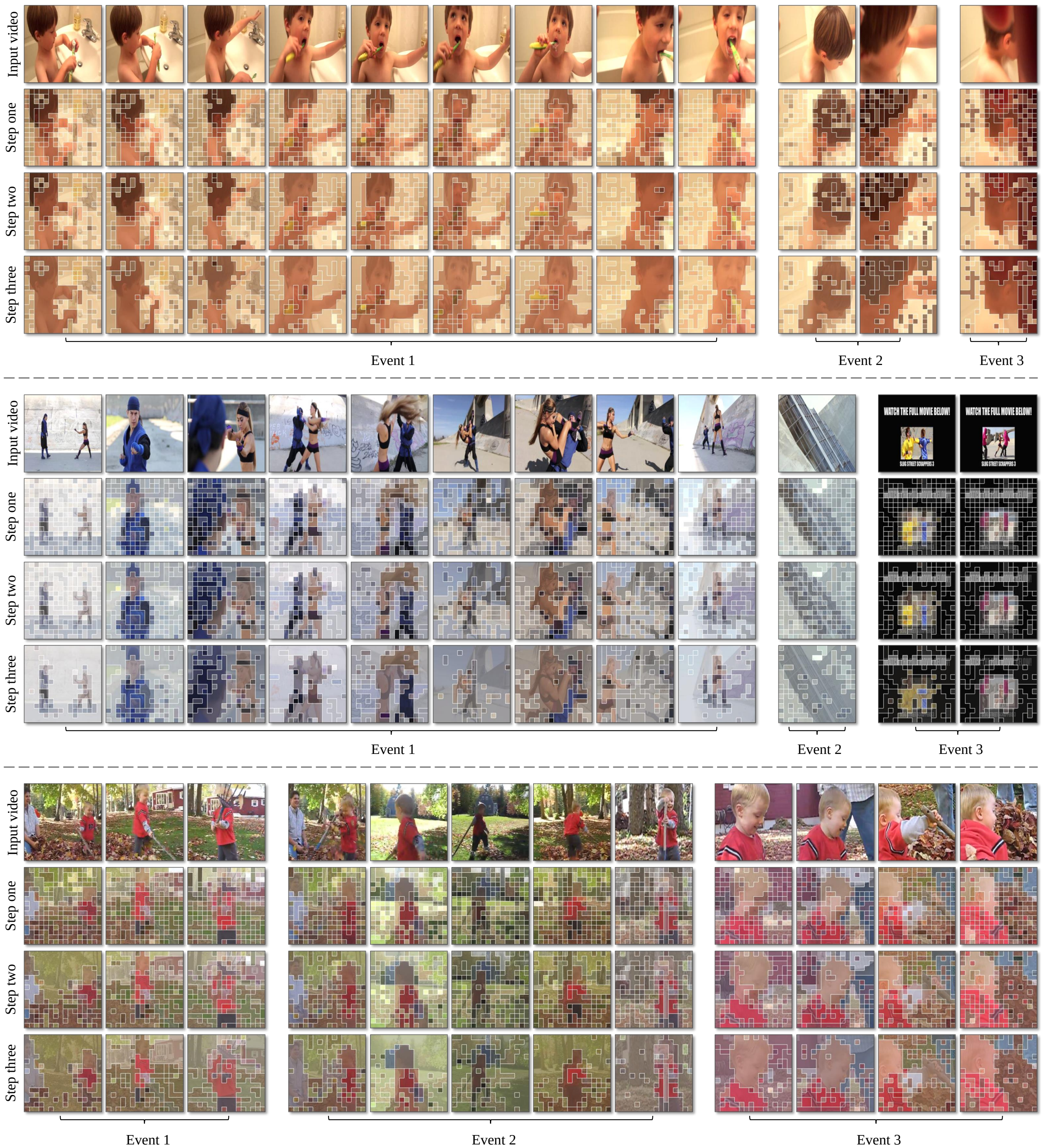


Figure B. **Visualization of the dynamic visual tokens for the video inputs.** It is important to emphasize that our proposed token merging method is parameter-free and operates without the need for object outline labels. Our method imposes no restrictions on the number of frames per event, showcasing the remarkable flexibility and generalization ability of our methodology.

tably, our method imposes no restrictions on the number of frames per event, showcasing the remarkable flexibility and generalization ability of our methodology. As shown in Fig. B, the proposed dynamic visual tokens significantly reduce the number of visual tokens while maintaining the expressive capabilities of the model. This empowerment equips Chat-UniVi with the capacity to capture the broader temporal understanding required for videos, all within the confines of a limited number of visual tokens.

D.3. Examples of Conversations

The conversation includes both the image and the video. In Fig. C and Fig. D, we present examples of conversations that encompass both the image and the video. As shown in Fig. C and Fig. D, Chat-UniVi offers detailed and contextually appropriate responses aligned with user prompts. These illustrative examples showcase the remarkable ability of Chat-UniVi to comprehend both image and video contexts across multiple conversational turns.

The conversation includes multiple videos. Fig. E illustrates a conversation example including multiple videos. As shown in Fig. E, Chat-UniVi can use the information of multiple videos in the context, and provide appropriate and coherent responses based on user prompts. The illustrative example showcases the remarkable ability of Chat-UniVi to comprehend multiple video contexts across multiple conversational turns.

The conversation includes multiple images. Fig. F provides an illustrative conversation example including multiple images. As shown in Fig. F, Chat-UniVi adeptly leverages information from multiple images within the context, enabling it to make choices among various images. This illustrative example highlights the impressive capacity of Chat-UniVi to grasp multiple image contexts seamlessly throughout various conversational exchanges.

The conversation includes the image. Fig. G features an example of a conversation that incorporates an image. As shown in Fig. G, Chat-UniVi excels at providing detailed descriptions and can even craft compelling narratives inspired by the image. The illustrative example showcases the remarkable ability of Chat-UniVi in the realms of reasoning and creative expression.

The conversation includes the video. In Fig. H and Fig. I, we offer examples of conversations that incorporate the video. As shown in Fig. H and Fig. I, Chat-UniVi exhibits a remarkable proficiency in comprehending videos and is adept at offering valuable insights inspired by the video content. These illustrative examples showcase the remarkable ability of Chat-UniVi to grasp video contexts and engage in reasoned responses.

E. Details of Quantitative Evaluations

E.1. GPT-based Evaluation for Image Understanding

Our quantitative evaluation protocol follows that of Liu et al. [16]. Following Liu et al. [16], Zhang et al. [30], we employ 90 questions based on 30 COCO validation images, covering various aspects, including conversation, detail description (Detail), and complex reasoning (Reason). These images are randomly selected by Liu et al. [16]. We utilize the GPT-4 model to generate reference responses based on the question, and the ground-truth bounding boxes and captions. During the model evaluation process, the model predicts answers based on both the question and input image. After obtaining the response from the model, we feed the question, visual information (in the format of captions and bounding boxes), the generated response, and the reference response to GPT-4. GPT-4 evaluates the helpfulness, relevance, accuracy, and level of detail of the responses, assigning an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Besides, we also ask GPT-4 to provide a comprehensive explanation of the evaluation to enhance our understanding of the models.

E.2. GPT-based Evaluation for Video Understanding

The quantitative evaluation protocol for video understanding follows the methodology introduced by Maaz et al. [18]. Specifically, Maaz et al. [18] curates a test set based on the ActivityNet-200 dataset [5], which includes videos with rich, dense descriptive captions and associated question-answer pairs from human annotations. During the model evaluation process, we employ the GPT-3.5 model to assign a relative score to the generated predictions on a scale of 1-5, across five critical aspects: (1) Correctness of information (Correct). (2) Detail orientation (Detail). (3) Contextual understanding (Context). (4) Temporal understanding (Temporal). (5) Consistency. It is worth noting that the results reported in Maaz et al. [18] span a range from 0 to 5. To standardize the metrics, we normalize all scores to a scale of 0 to 100.

E.3. Zero-shot Video Question Evaluation

Our evaluation protocol follows that of Maaz et al. [18], utilizing GPT-assisted evaluation to assess the capabilities of models. During the model evaluation process, we feed the question, the ground-truth answer, and the generated response to the GPT-3.5 model. GPT-3.5 evaluates whether the generated responses are correct and assigns a matching score on a scale of 0 to 5, where a higher score indicates better overall performance.

E.4. Zero-shot Object Hallucination Evaluation

To quantitatively evaluate the hallucination problem of the model, we adopt the polling-based object probing evaluation (POPE) process proposed by Li et al. [13]. Specifically, POPE formulates the evaluation of object hallucination as a binary classification task, where the model is prompted to respond with either "Yes" or "No" to queries like "Is there a chair in the image?". Li et al. [13] randomly selects 500 images from the COCO validation set. Each image contains more than three ground-truth objects in the annotations, and six questions are generated for each image. The annotations of objects in images directly construct the questions with the answer "Yes". For the questions with the answer "No", three different strategies are employed for sampling their probing objects as follows:

- **Random Sampling.** Randomly sampling objects that do not exist in the image.
- **Popular Sampling.** Selecting the top-3 most frequently occurring objects in the COCO dataset that are absent from the image.
- **Adversarial Sampling.** Initially, Li et al. [13] rank all objects based on their co-occurring frequencies with the ground-truth objects, and subsequently select the top-3 most frequent objects from this list that are not present in the image.

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, 2

[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. 4

[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. 1

[4] 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. 3

[5] 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. 9

[6] 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. 1

[7] 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. 4

[8] 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. 1

[9] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 5

[10] 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. 4

[11] 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. 2

[12] 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. 1, 4

[13] 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, 10

[14] 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. 1

[15] 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. 1

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 4, 9

[17] Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points. In *ICLR*, 2023. 2

[18] 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. 4, 9

[19] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022. 3

[20] 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

Tell me about the movie.

The movie is a computer-animated superhero film featuring the iconic character Spiderman. It is set in a bustling city where Spiderman, a superhero with incredible spider-like abilities, fights crime and protects the innocent. The film showcases Spiderman's adventures as he navigates the urban landscape, using his spider-sense and web-shooting abilities to take down villains and save lives. The movie captures the essence of the Spiderman character, combining action, adventure, and humor, while also exploring themes of responsibility and heroism.

Write a rich narration for this video based on the background of the movie.

In a bustling metropolis, a man dressed in a red and black Spiderman suit, clutching a backpack, swings through the air, defying gravity. He soars above the busy city streets, where a multitude of vehicles, including cars, trucks, and buses, navigate through the chaos. The Spiderman suit stands out against the urban backdrop, a symbol of hope and heroism in a world of darkness. The city's towering buildings and neon signs create a dazzling backdrop for Spiderman's daring feats, as he leaps from rooftop to rooftop, taking on the city's criminal underworld. The city's inhabitants look up in awe and wonder at the spectacle unfolding before their eyes, as the Spiderman legend continues to inspire and captivate generations to come.

If Spiderman and Batman were to make a movie together, what do you think the story would be?

If Spiderman and Batman were to make a movie together, it would be a thrilling and action-packed adventure that would showcase their unique abilities and teamwork. The story could revolve around a common enemy, such as a powerful villain who poses a threat to both the city of New York and Gotham City. Spiderman and Batman would have to work together to uncover the villain's plans and put an end to their evil schemes.

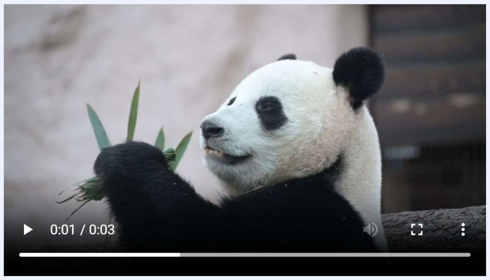
The story could begin with Spiderman swinging through the streets of New York, trying to track down a lead on the villain. Meanwhile, in Gotham City, Batman is investigating a series of strange occurrences that seem to be connected to the same villain. As they both close in on the villain, they discover that they are working together and must put aside their differences to stop them.

Throughout the movie, we could see the dynamic between Spiderman and Batman, highlighting their unique approaches to fighting crime and their contrasting personalities. Spiderman, being a web-slinging hero, would rely on his agility and quick reflexes, while Batman would use his detective skills and martial arts prowess to take down the villain.

The climax of the movie would involve a thrilling showdown between Spiderman and Batman and the villain, with both heroes putting their lives on the line to save the city. In the end, they would emerge victorious, having not only saved the day but also forged a newfound respect and admiration for each other's abilities. The movie would conclude with a final shot of the two heroes, standing together, ready to protect the city from any threat that may come their way.

Figure C. **A conversation with both image and video.** The blue box shows the user input. The gray box shows the model output.

Can you write a short poem inspired by this video?



Sure, here's a short poem inspired by the video:

A panda bear in a zoo,
Eating bamboo with a smile anew.
It reaches up with a paw,
To grab some greens to munch on now.

The zookeeper feeds it bamboo,
And the panda bear enjoys it so.
It chews and grinds the plant,
And savors every single crumb.

A panda bear in a zoo,
A simple life, yet so true.
It eats its greens with glee,
And finds joy in the small things, you see.

If I want to use this image for the zoo's publicity, do you have any suggestions for designing the poster?

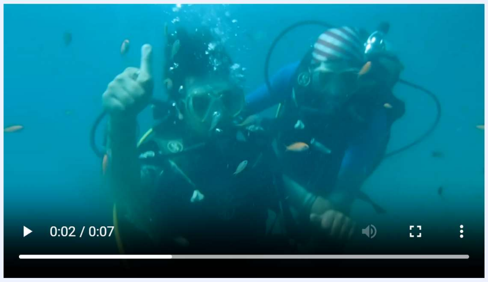
Yes, I have a few suggestions for designing a poster using this image. Since the image features a panda bear wearing a hat, you could use this unique and eye-catching visual to create a memorable and playful poster. Here are a few ideas:

1. Keep it simple: Use a clean and minimalist design to highlight the panda bear's adorable appearance and the hat. You could use a large photo of the panda bear with the hat in the center of the poster, and keep the text to a minimum.
2. Add a tagline: Include a short and catchy tagline, such as "Experience the cuteness overload" or "Visit the zoo and meet the cutest animal," to capture the viewer's attention and encourage them to visit the zoo.
3. Use bold colors: Choose bold and vibrant colors to make the poster stand out, such as using the panda bear's natural colors as a base and adding contrasting colors for the hat and background.
4. Add more visuals: To make the poster more dynamic, you could include additional visuals, such as other animals wearing hats or interacting with the panda bear, fun facts about pandas, or images of zoo visitors enjoying their time with the panda bear.
5. Include a call-to-action: Encourage viewers to visit the zoo by including a clear call-to-action, such as "Visit us today to see this adorable panda bear in person" or "Share this poster with your friends and family to spread the cuteness."

By incorporating these suggestions, you can create an eye-catching and memorable poster that showcases the panda bear's unique appearance and encourages zoo visitors to experience the cuteness for themselves.

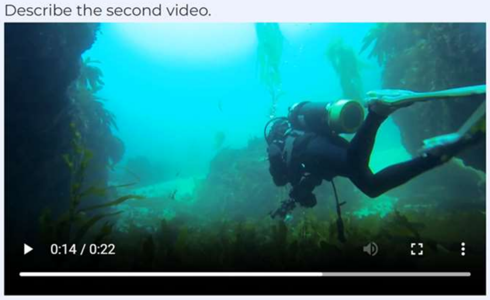
Figure D. **A conversation with both image and video.** The blue box shows the user input. The gray box shows the model output.

Describe the video in as much detail as possible.



The video shows two scuba divers underwater, exploring the beautiful and diverse ecosystem of the ocean. They are seen swimming around, observing the marine life and admiring the coral reefs. The scuba divers are wearing proper diving gear, including oxygen tanks, masks, and wetsuits. The water is crystal clear, allowing the divers to see the vibrant colors of the fish and other marine creatures. The coral reefs are home to a variety of species, including colorful fish, sea turtles, and sea anemones. The divers move slowly and deliberately, ensuring they do not disturb the delicate ecosystem. The video captures the peaceful and serene atmosphere of the underwater world, with the sound of the divers' bubbles and the gentle lapping of the waves providing a soothing background sound. Overall, the video offers a mesmerizing glimpse into the beauty and complexity of the ocean's ecosystem, and the importance of preserving it for future generations.

Describe the second video.



The second video shows a group of scuba divers exploring the ocean floor. They are seen swimming around and observing the marine life. The video captures the divers' movements as they move around the ocean floor, and the camera follows them closely, providing a clear view of the underwater environment. The divers are equipped with scuba gear, and they appear to be in awe of the vibrant and diverse marine life surrounding them. The video showcases the beauty of the underwater world and the importance of preserving it.

If I want to splice these two videos together, do you have any suggestions?

Yes, I can provide some suggestions for splicing these two videos together. First, you can start with the first video and show the scuba divers swimming around the coral reef and observing the marine life. Then, you can transition to the second video and show the divers swimming around the ocean floor and observing the diverse marine life. You can use a smooth transition effect to blend the two videos seamlessly. After that, you can transition back to the first video and show the divers swimming around the coral reef and observing the marine life. Finally, you can end the video with a final shot of the divers swimming away from the camera, leaving the viewers with a sense of awe and appreciation for the beauty of the underwater world.

Figure E. **A conversation includes multiple videos.** The blue box shows the user input. The gray box shows the model output.

Describe the two images separately.

The first image shows a large brown bear sitting on a grassy field, while the second image shows a woman leaning on a large brown teddy bear.

How to make the bear in the first image as cute as the bear in the second image?

To make the bear in the first image as cute as the bear in the second image, you can try the following:

1. Position the bear in a more endearing pose, such as sitting or lying down in a comfortable and relaxed manner, instead of standing upright.
2. Add some accessories to the bear, such as a cute little hat, bow tie, or even a pair of glasses, to give it a more human-like appearance.
3. Use a soft, pastel-colored background to create a calming and gentle atmosphere, which can make the bear appear even more adorable.
4. Adjust the lighting to create a warm and inviting ambiance, as this can also enhance the bear's cuteness.
5. Experiment with different camera angles and perspectives to capture the bear's features and expressions in detail, making it appear more lifelike and endearing.

By incorporating these suggestions, you can make the bear in the first image appear as cute as the bear in the second image.

Which of these two bow ties fits the bear better?

The pink satin bow tie fits the teddy bear better than the black bow tie.

Give your reasons.

The pink satin bow tie is larger and more visible on the teddy bear, making it a more prominent feature. The black bow tie is smaller and less noticeable, which might make the bear look less attractive. Additionally, the pink color of the bow tie adds a pop of color that contrasts with the brown fur of the bear, making it look more adorable and appealing.

Figure F. **A conversation includes multiple images.** The blue box shows the user input. The gray box shows the model output.

How many superheroes are there in this image?

There are two superheroes in this image.

Describe each of them in detail.

In the image, there are two superheroes engaged in an action-packed fight. One of them is wearing a red and yellow suit, which is a classic representation of the iconic superhero Iron Man. Iron Man is known for his high-tech armor and powerful weapons, as well as his intelligence and wit. He is a founding member of the Avengers and has been portrayed in various media, including movies and comic books.

The other superhero is wearing a blue and red costume, which is reminiscent of the character Captain America. Captain America is a symbol of patriotism and is known for his exceptional strength, endurance, and moral values. He is also a founding member of the Avengers and has been portrayed as a key figure in various Marvel Comics storylines.

Both superheroes are depicted in mid-air, leaping towards each other with great force, creating an intense and exciting scene. Their clash highlights their incredible abilities and showcases their epic battle in the comic book or movie from which this image is taken.

Write a movie story based on this image.

The movie opens with a dramatic scene of the two superheroes, Iron Man and Captain America, leaping towards each other in mid-air. The camera captures their intense expressions and the determination in their eyes as they engage in a fierce battle. The audience is left wondering about the reason behind their fight and what could be at stake.

As the story unfolds, we learn that the conflict between Iron Man and Captain America stems from a disagreement on how to handle a dangerous threat that has emerged. Iron Man believes that the use of force is necessary to protect humanity, while Captain America holds a more idealistic view and seeks to resolve the situation through peaceful means.

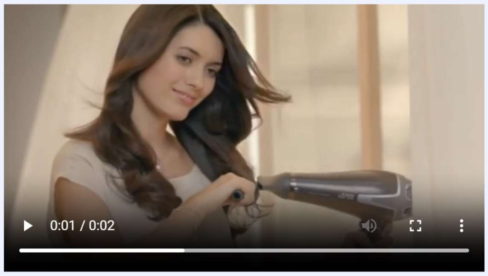
Their battle takes them across various locations, from the streets of a bustling city to the depths of a secret underground facility. Along the way, they encounter other superheroes who are caught in the crossfire and must choose sides. The tension and drama continue to build as the story progresses, with each character revealing their motivations and backstory.

As the climax approaches, the two superheroes come face to face in an epic showdown. They engage in a thrilling fight, showcasing their unique abilities and skills. In the end, they realize that their conflict has brought them closer to the truth and that their differences have made them stronger. They put aside their differences and unite to face the common enemy, saving the world and restoring peace.

The movie concludes with a final scene that highlights the importance of teamwork and the power of unity. The superheroes stand together, ready to face any challenge that comes their way, while the audience is left with a sense of awe and admiration for these extraordinary characters.

Figure G. **A conversation includes the image.** The blue box shows the user input. The gray box shows the model output.

What is the video about?



The video is about a woman who is blow drying her hair in a room.

Generate a creative advertising feature about the hair dryer used by the girl, which can be used on the product website.

The hair dryer used by the girl in the video is not just any ordinary hair dryer, it's a smart hair dryer that comes with a built-in intelligence. This smart hair dryer uses advanced technology to analyze the hair type and drying conditions, ensuring that the hair is dried evenly and efficiently. The smart hair dryer also comes with a built-in temperature and airflow control system, allowing the user to customize the drying experience to their hair type.

With the help of the smart hair dryer, the girl in the video was able to achieve a sleek and polished look in just a few minutes. The hair dryer is also lightweight and easy to use, making it a great option for busy women on the go.

In addition to its advanced features, the smart hair dryer is also environmentally friendly. It uses less energy and produces less heat than traditional hair dryers, making it a more sustainable option.

To learn more about this amazing hair dryer and its features, visit the product website. With its advanced technology and eco-friendly design, this smart hair dryer is the perfect tool for any woman looking to achieve a sleek and polished look.

If this video is a commercial for this hair dryer, do you have any suggestions to make this video more popular?

To make this video more popular, we can consider the following suggestions:

1. Include before-and-after shots: Showing the difference in the girl's hair before and after using the hair dryer can help viewers understand the effectiveness of the product.
2. Add testimonials: Including testimonials from satisfied customers who have used the hair dryer can help build trust and credibility for the product.
3. Highlight unique features: Emphasize the unique features of the hair dryer, such as the smart technology, temperature control, and airflow adjustment, to showcase its advantages over other hair dryers.
4. Showcase different hair types and styles: Demonstrate how the hair dryer works on different hair types and styles, showcasing its versatility and adaptability.
5. Include a call-to-action: Encourage viewers to visit the product website or make a purchase by including a clear call-to-action at the end of the video.
6. Optimize for social media: Create short clips or teasers from the video to share on social media platforms, driving traffic to the product website and increasing brand awareness.

By incorporating these suggestions, the video can become more engaging, informative, and persuasive, ultimately leading to increased interest and sales of the hair dryer.

Figure H. **A conversation includes the video.** The blue box shows the user input. The gray box shows the model output.

Explain why this video is funny.

The video is funny because it features a person wearing a large, humorous teddy bear costume. The individual's choice to wear this oversized and comical outfit creates a sense of amusement and playfulness, which is amplified when they begin to dance. The combination of the exaggerated size of the costume and the person's decision to dance adds a lighthearted and entertaining element to the video. This unconventional and whimsical scene captures the viewer's attention and evokes a sense of joy and laughter, making it an enjoyable and humorous experience.

Do you have any suggestions to make this video more popular?

To make this video more popular, several strategies could be employed. Here are a few suggestions:

1. Social Media Sharing: Encourage the person in the teddy bear costume to share the video on their social media accounts, tagging friends and family to help expand its reach. Utilizing hashtags relevant to the content can also increase visibility to a broader audience.
2. Collaborations: Collaborate with other content creators, influencers, or popular personalities to feature the video or create a crossover content. This can help attract their followers and expose the video to a larger audience.
3. Paid Advertising: Invest in targeted online ads or promotions, such as Google AdWords or social media ads, to drive traffic to the video. This can be an effective way to reach a specific demographic or location, increasing the video's visibility and potential viewership.
4. Optimize Video Content: Ensure that the video is well-edited, visually appealing, and includes engaging elements such as captions, special effects, or additional footage. This can help capture the viewer's attention and keep them engaged throughout the video.
5. Engage with the Audience: Encourage viewers to like, comment, and share the video. Responding to comments and interacting with the audience can help create a sense of community and increase the video's popularity by organically spreading the content through word-of-mouth.

By implementing these strategies, the video's popularity can be increased, potentially leading to more views, shares, and engagement with the content.

Figure I. **A conversation includes the video.** The blue box shows the user input. The gray box shows the model output.

[21] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 3

[22] 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. 1

[23] 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. 4

[24] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 5

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

[26] 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. 4

[27] 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. 1

[28] 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. 2

[29] 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. 2

[30] 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. 9

[31] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, pages 12697–12706, 2021. 3

[32] 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. 1