Towards Language-Driven Video Inpainting via Multimodal Large Language Models

Jianzong Wu[1,3]* Xiangtai Li[2,3] † Chenyang Si[2] Shangchen Zhou[2] Jingkang Yang[2]
Jiangning Zhang[5] Yining Li[3] Kai Chen[3] Yunhai Tong[1,4] Ziwei Liu[2] Chen Change Loy[2]
[1] National Key Laboratory of General Artificial Intelligence, Peking University [2]S-Lab, Nanyang Technological University
[3]Shanghai AI Laboratory [4] PKU-Wuhan Institute for Artificial Intelligence [5] Zhejiang University
Project Page: https://jianzongwu.github.io/projects/rovi

Abstract

We introduce a new task – language-driven video inpainting, which uses natural language instructions to guide the inpainting process. This approach overcomes the limitations of traditional video inpainting methods that depend on manually labeled binary masks, a process often tedious and labor-intensive. We present the Remove Objects from Videos by Instructions (ROVI) dataset, containing 5,650 videos and 9,091 inpainting results, to support training and evaluation for this task. We also propose a novel diffusion-based language-driven video inpainting framework, the first end-to-end baseline for this task, integrating Multimodal Large Language Models to understand and execute complex language-based inpainting requests effectively. Our comprehensive results showcase the dataset's versatility and the model's effectiveness in various language-instructed inpainting scenarios. We have made datasets, code, and models publicly available at https://github.com/jianzongwu/Language-Driven-Video-Inpainting.

1. Introduction

Video inpainting, a technique for restoring missing or corrupted segments in video frames, finds extensive application in areas such as video completion [4], video restoration [15], and object removal [6]. Despite continuous advancements in enhancing image quality and temporal coherence of inpainting results [10, 23, 47, 69], current methods predominantly depend on *manually annotated binary masks* to identify restoration areas. This manual process is time-consuming and impractical for long videos. While automatic labeling tools, such as segmentation and object tracking models [57, 61, 64], offer some relief, they often necessitate manual refinement due to imperfect labeling.

*The work is done during the internship at Shanghai AI Laboratory. †: Project Lead and corresponding author. Email: xiangtai94@gmail.com

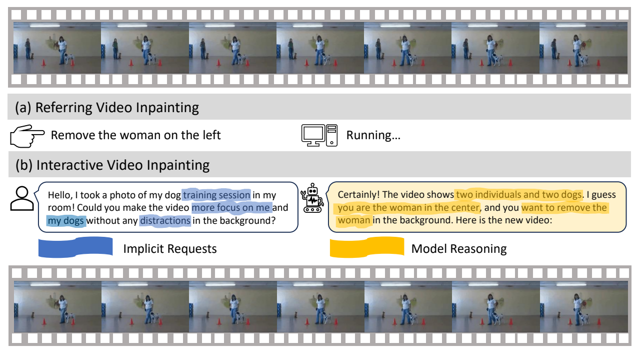


Figure 1. **Language-driven video inpainting.** It contains two sub-tasks based on the expression types. The referring video inpainting task takes simple referring expressions as input, while interactive video inpainting receives chat-style conversations. The conversation may encounter implicit requests, and the model needs to reason for a correct understanding.

Perhaps a more natural way to perform video inpainting is through natural language, as shown in Fig. 1. The task would become much easier if we could leverage natural language descriptions to specify the inpainting areas, like "woman on the left," thereby preventing the need for pixel-level manual annotations. Importantly, the language-driven setting can benefit from the flexibility of natural language. For example, with richer sentences, one can easily refer to multiple or abstract objects, which is much more effective than labeling masks. Extending from this notion, one could divide the task into two subtasks, namely "Referring Video Inpainting" and "Interactive Video Inpainting." The former takes simple referring expressions as inputs, and the latter considers more complex conversation-like interactions to accomplish the inpainting task.

To establish a model for the proposed tasks, it is essential to have an appropriate dataset for both training and evaluation. Currently, no publicly available dataset comprises the triplets of original videos, removal expressions, and inpainted videos. In response to this gap, we build a new

Figure 2. **Comparison with general image editing models**. InstructPix2Pix [1] and MagicBrush [66] are general image editing methods based on diffusion models. They produce inferior results when instructed to remove objects in videos.

dataset named the Remove Objects from Videos by Instructions (ROVI) dataset. Specifically, we employ referring object segmentation datasets, which are pre-annotated with object masks and descriptive expressions. These datasets are further augmented with corresponding inpainted videos generated using a state-of-the-art video inpainting model. However, we find existing referring expressions for interactive video inpainting tasks too simplistic. To address this limitation, we employ Multimodal Large Language Models (MLLMs) [3, 63, 70] to create conversation-like dialogues. These dialogues are designed to simulate real-world scenarios, encompassing user requests and corresponding machine responses. This approach enriches the dataset, making it more representative of the complexity and variability found in practical video inpainting applications.

In addition to the dataset, we introduce the first end-to-end baseline model, Language-Driven Video Inpainting (LGVI), for the proposed tasks. Our model is built upon diffusion-based generative models. In particular, we inflate the text-to-image (T2I) model to become a text-to-video (T2V) architecture by extending the parameters with an additional temporal dimension. We propose an efficient visual conditioning approach that minimally increases the number of parameters. To further enhance our model's capabilities for the interactive task, we extend the LGVI framework to LGVI-I (Interactive). This extension incorporates an MLLM specifically designed to process and understand user requests phrased in a conversation-like format. The LGVI-I model is trained in an end-to-end manner. This interactive architecture enables the system to interpret complex instructions accurately. As a result, it can produce appropriate inpainting results and relevant responses within the conversational context, thus paving the way for more intuitive and flexible user interactions with interactive video inpainting systems.

In summary, our key contributions are as follows:

- We introduce a novel language-driven video inpainting task, significantly reducing reliance on human-labeled masks in video inpainting applications. This task includes two distinct sub-tasks: referring video inpainting and interactive video inpainting.
- We propose a dataset to facilitate training and evaluation for the proposed tasks. This dataset is the first of its kind, containing triplets of original videos, removal expressions, and inpainted videos, offering a unique resource for research in this domain.
- We present a diffusion-based architecture, LGVI, as a baseline model for the proposed task. We show how one could leverage MLLMs to improve language guidance for interactive video inpainting. To our knowledge, it is the first model to perform end-to-end language-driven video inpainting.

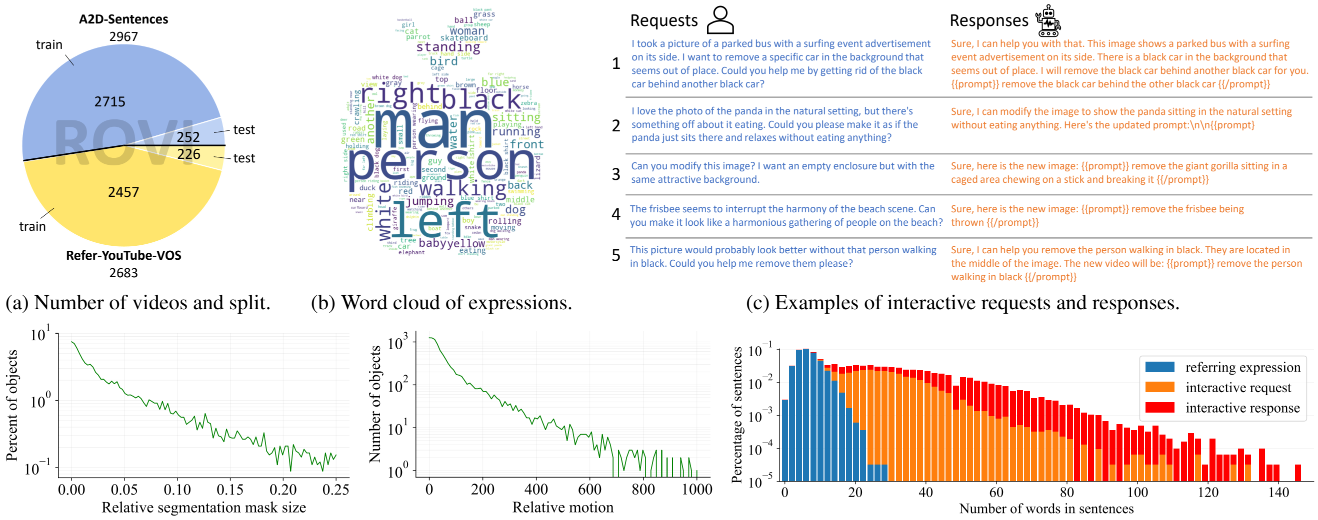
2. Related Work

Video inpainting. Video inpainting is a technique aimed at restoring or filling missing or corrupted parts in a video plausibly. While related to image inpainting methods [18, 19, 24, 25, 36, 60], video inpainting techniques [4, 10, 16, 23, 27, 47, 65, 69] extend the problem to the more complex domain of moving visuals. This technique can be applied for various applications, such as object removal, visual restoration, and completion. With the advent of deep learning, visual inpainting networks usually employ convolutional neural networks (CNNs) [10, 47, 60] and generative adversarial networks (GANs) [18, 24, 25, 36]. Recent works also apply vision Transformers [5, 30, 33, 40, 59] to enhance the global interaction among visual features [16, 19, 23, 65, 69]. State-of-the-art methods show strong abilities in restoring missing parts and removing objects. Most of these works require the input of a binary mask to define the restoring area [4, 19, 23, 24, 27, 65]. However, the generation of object-like masks, particularly for lengthy videos, poses a significant and labor-intensive challenge,

Language-driven visual editing. Diffusion-based text-to-image generation models (DMs) [34, 39, 41, 42, 44] show excellent abilities in generating images and videos following text guidance. Recent studies also achieve image editing [1, 9, 32, 45, 66], image segmentation and grouping [20–22, 50–52, 56, 62, 68] and video editing [12, 53] with DMs. Among these, Prompt2Prompt [9] manipulates the cross-attention maps within the model to enable various editing operations such as object modification, addition, and style transfer. InstructPix2Pix [1] leverages this approach to create an image editing dataset. Similarly, Tune-A-Video [53] proposes a training-free architecture to edit videos by language references. However, these works are intended for general visual editing. They tend to yield suboptimal results when applied to more specific challenges, such as language-driven video inpainting. Figure 2 shows two examples where these models produce inferior results when instructed to remove objects. A few works have explored the image inpainting task using DMs. Repaint [31]

Table 1. **Comparison between the ROVI dataset and related datasets.** We choose two commonly used image inpainting (II) datasets, two video inpainting (VI) datasets, and one language-guided image inpainting (LII) dataset. Our ROVI dataset is the first for language-guided video inpainting (LVI) and interactive video inpainting (IVI) tasks.

Dataset	Task	Scene	#Images	#Videos	#Frames	Annotations				#Objects	#Exprs	#Chats	Source
						mask	expr	inpaint	chat				
Places [67]	II	Buildings & Places	10,624,928	-	-	×	×	×	×	-	-	-	-
CelebA [29]	II	Human Face	202,599	-	-	×	×	×	×	-	-	-	-
YouTube-VOS [55]	VI	General	-	4,453	197,272	✓	×	×	×	7,755	-	-	YouTube
DAVIS [37]	VI	General	-	50	3,455	×	×	×	×	3,455	-	-	-
GQA-Inpaint [58]	LII	General	49,311	-	-	✓	✓	✓	×	97,854	107,252	-	GQA [11]
ROVI	LVI + IVI	General	-	5,650	247,018	✓	✓	✓	✓	9,091	12,534	9,091	Refer-YouTube-VOS [43] + A2D-Sentences [7]



(a) Number of videos and split. (b) Word cloud of expressions. (c) Examples of interactive requests and responses.

(d) Relative segmentation mask size. (e) Relative object motion distribution. (f) Sentence lengths distribution.

Figure 3. **The ROVI dataset statistics.** Best viewed in color.

takes the image and mask as input and lets the DM restore the original image. SmartBrush [54] takes mask and text as input to guide a region-controlled generation on the masked area, which aims to generate new concepts rather than remove the object. Recently, Inst-Inpaint [58] proposes a method to perform object removal on images based on the language descriptions. Despite its innovative approach, Inst-Inpaint's training samples are constrained by a lack of interactive expressions and video resources, which limits its practical effectiveness in complex scenarios.

Multi-Modal Large Language Models. Large language models (LLMs) have demonstrated exceptional performance across a variety of text-based tasks and applications [2, 3, 8, 17, 35, 46]. Recent works extend the capabilities of LLMs to include image processing and computer vision. A notable example is LLaVA [26], which translates image tokens into a language feature space, thereby transforming the fine-tuned model into an MLLM. This adaptation enables MLLMs to interpret and understand visual content. Subsequent research uses MLLMs in diverse applications, including visual reasoning, object detection, and segmentation [8, 17, 26, 38, 63, 70]. To the best of our knowledge, this study is the first to integrate MLLMs into the domain of language-driven video inpainting.

3. ROVI Dataset

3.1. Comparing with Existing Datasets

Table 1 summarizes the differences between ROVI and related datasets. In image and video inpainting research, prevalent training and evaluation datasets mainly include vision-centric ones like Places [67], CelebA [29], YouTube-VOS [55], and DAVIS [37], without human annotations. These datasets typically employ random masking in training to simulate missing areas for inpainting. However, for object removal tasks, specifically labeled masks are essential. While YouTube-VOS provides object masks, it lacks corresponding inpainting ground truths. The GQA-Inpaint dataset [58], although rich in object expressions and inpainting results, is limited to image data and does not accommodate video or interactive contexts. Our ROVI dataset addresses these limitations with comprehensive annotations covering a wide array of regions, including object masks, referring expressions, inpainting results, and conversation-like dialogues. Unlike Places and CelebA, which focus on specific image categories like buildings or faces, ROVI en-

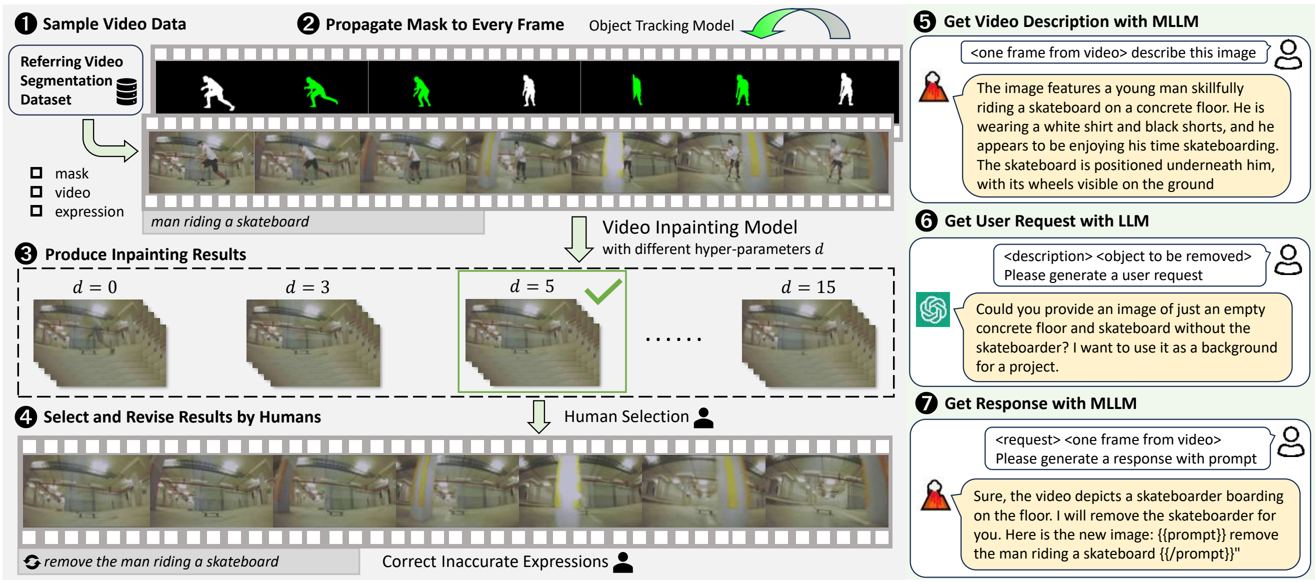


Figure 4. **ROVI dataset annotation pipeline**. The building process of the ROVI dataset involves two distinct phases: inpainting annotation and interactive annotation. In the inpainting annotation phase, the primary objective is to incorporate inpainting results into existing referring video segmentation datasets, which initially contain object masks and expressions. During the interactive annotation pipeline, we follow a multi-step approach incorporating LLMs and MLLMs. Best viewed in color.

compasses a broader spectrum of general scenes, making it more adaptable for diverse inpainting applications.

3.2. Dataset Statistics

Figure 3 presents a comprehensive statistical analysis of the ROVI dataset. The dataset encompasses 2,967 videos from A2D-Sentences and 2,683 videos from Refer-YouTube-VOS, divided into train and test splits, as shown in Fig. 3a. Figure 3b illustrates the diversity of referring expressions with word clouds. Figure 3c shows several examples of our dataset's interactive requests and responses, showing the diversity and complexity of the dialogues. Figure 3d details the relative sizes of segmentation masks (mask area divided by image area). We drop objects with a relative size larger than 0.25 because the inpainting results for large objects usually have worse qualities. Figure 3e analyzes the distribution of object motion. Figure 3f delivers a histogram of sentence lengths within the dataset.

3.3. Dataset Construction Pipeline

Video data selection. As depicted in Fig. 4, we have chosen referring video object segmentation datasets for the source of video data. Referring video object segmentation aims to segment an object referred to by a given language description. These datasets have pre-annotated object masks and descriptive expressions, making them well-suited for the proposed task. Specifically, we select Refer-YouTube-VOS [43] and A2D-Sentences [7] as our data sources.

Annotation pipeline. We use a video inpainting model to generate the inpainting ground truth. Specifically, we choose E^{2}FGVI [23], a state-of-the-art video inpainting model, to produce the inpainting results. This model, trained on video data, guarantees temporal consistency in the inpainting results.

To further ensure the ground truth is of high quality, we incorporate a human selection process on the hyperparameter of the inpainting method. Specifically, the input mask can be expanded with different pixel sizes, denoted as d. The bigger the d is, the larger the input mask is developed so that it may cover the whole object. The best d value varies through objects, causing an unstable performance in the inpainted videos if set to a fixed value. Therefore, throughout the data generation process, we experiment with various hyperparameters to generate multiple results for each object and involve human annotators to select the best result. More details are provided in the supplementary.

For interactive annotations, we need to collect expressions through chat-style conversations. Unlike the straightforward "remove" sentences, these interactive requests should be implicit, necessitating the model to discern the user's underlying intent. Rather than relying solely on human annotators to articulate these requests, we explore a more automated approach: we employ LLMs and MLLMs to simulate a human user and generate potential requests and responses. We propose a multi-step approach with details illustrated in Fig. 4. By employing this dual-faceted annotation pipeline, the ROVI dataset is enabled to handle complex user requests.

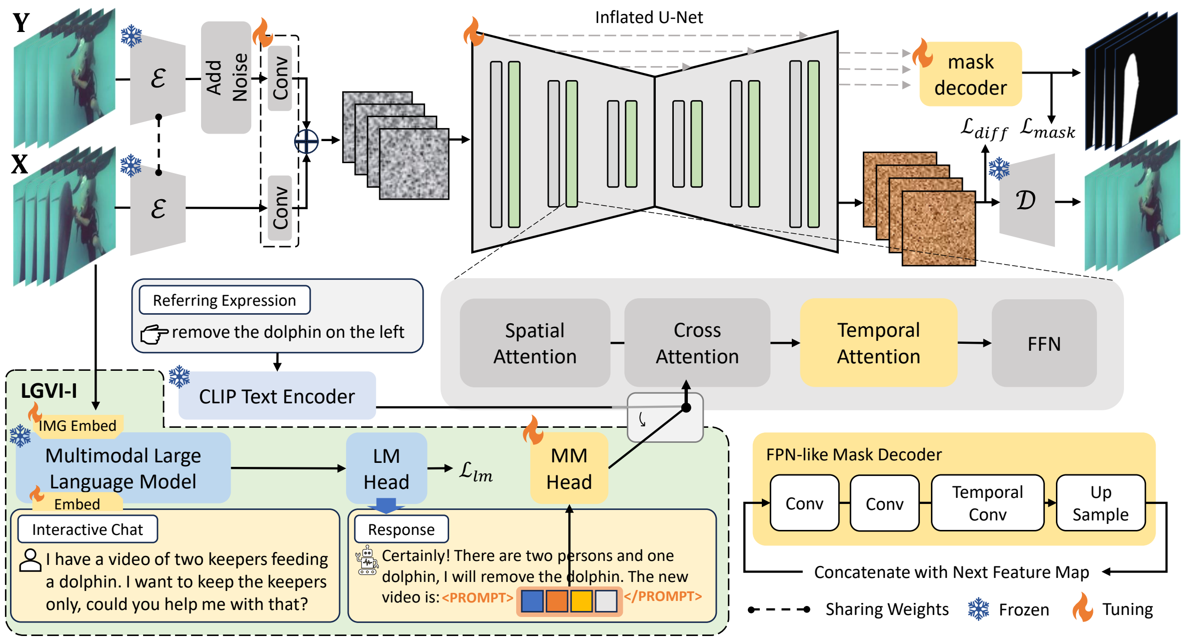


Figure 5. **The training process of LGVI and LGVI-I**. We inflate the U-Net with a temporal dimension to allow video input. To ensure temporal consistency in the generated videos, we introduce a temporal attention module between cross-attention and FFN layers. Additionally, we propose a mask decoder module for explicit guidance in inpainting tasks. We augment LGVI with MLLM joint training for interactive video inpainting, resulting in LGVI-I as the baseline. The output of MLLM includes a set of prompt tokens, which is fed into the cross attention of the U-Net.

4. Methodology

In this section, we introduce our Language-Driven Video Inpainting framework (LGVI) and the MLLM-enhanced LGVI-I (Interactive) architecture. The latter is built from the former architecture by adding extra LLM as language controllers.

4.1. LGVI

The LGVI framework is shown in Fig. 5, which is built on the architecture of Stable Diffusion [41]. To extend the framework to video inputs, we perform temporal inflation by reorganizing the network's structure following [44, 53]. For a batch video input with T frames, denoted as $\mathbf{X} \in \mathbb{R}^{B\times T\times H\times W\times 3}$, where B is the batch size, and $H \times W$ are the size, we transpose the tensor to $\mathbf{X} \in \mathbb{R}^{(B\times T)\times H\times W\times 3}$. This transformation converts the input into a 4-dimensional image batch input format. The pre-trained 2D networks can process video clips as they are separate images. Additionally, we introduce a parameter-efficient Temporal Attention module positioned between the cross-attention and FFN network. Given latent feature $\mathbf{v} \in \mathbb{R}^{(B\times T)\times D\times C}$, where D is the length of patched visual tokens, and C is the channel size, we transpose it to $\mathbf{v}' \in \mathbb{R}^{(B\times D)\times T\times C}$.

The Temporal Attention module is formulated as follows:

$$\mathbf{Q} = \mathbf{W}_q\mathbf{v}, \quad \mathbf{K} = \mathbf{W}_k\mathbf{v}, \quad \mathbf{V} = \mathbf{W}_v\mathbf{v},$$
$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}})\cdot\mathbf{V}, \tag{1}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable matrices to project the inputs to query, key, and value. The computational complexity of the Temporal Attention module is $\mathcal{O}(CT^2)$, while spatial self-attention has a complexity of $\mathcal{O}(CD^2)$. Considering $T \ll D$. The Temporal Attention module is a time-efficient tool to ensure the consistency of video sequences.

Diffusion models learn to gradually remove noises in a noised video. During training, the target video $\mathbf{Y}$ is added with noises to be the start point of the noised video. Besides the noised target video input, LGVI also takes the original video $\mathbf{X}$ as a control signal input. Concretely, we encode the original video $\mathbf{X}$ to the latent space and concatenate its feature with the noised target video in the channel dimension. Note that the noise is added only to the target video latent, and during testing, the noised target video is a randomly generated noise.

$$\begin{aligned}\mathbf{z}_0 &= \mathcal{E}(\mathbf{Y}), \quad \mathbf{c}_x = \mathcal{E}(\mathbf{X}),\\ \mathbf{z}_t &= \text{AddNoise}(\mathbf{z}_0),\\ \mathbf{v}_t &= \text{Conv}_v(\mathbf{z}_t) + \text{Conv}_x(\mathbf{c}_x),\end{aligned} \tag{2}$$

where $\mathcal{E}$ is the pre-trained VAE encoder, t is the sampled timestamp, Conv_v and Conv_x are convolutional layers with 3×3 kernels to transfer the latent codes into U-Net feature dimensions. The initial weights of Conv_x are set to all-zero. This technique allows the model to add video condition guidance during training. Due to mask annotations in the ROVI dataset, we can leverage masks as an additional supervision signal in our LGVI framework. Concretely, we implement a mask decoder to predict the object's mask in the video that needs to be inpainted or removed. This decoder uses the outputs from U-Net up-blocks and consists of convolutional and temporal convolutional layers. The use of mask supervision enables the model to focus on the region described in the natural language input, thereby facilitating precise and targeted inpainting. The effectiveness of mask supervision can be seen in Sec. 5. The training objective of LGVI is:

$$\begin{aligned} \mathcal{L}_{diff} &= \mathbb{E}_{\mathbf{X},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[||\epsilon - \epsilon_\theta(\mathbf{v}_t, \mathbf{c}, t)||_2^2\right], \\ \mathcal{L}_{mask} &= \mathrm{CrossEntropyLoss}(\hat{\mathbf{M}}, \mathbf{M}), \\ \mathcal{L} &= \lambda_1\mathcal{L}_{diff} + \lambda_2\mathcal{L}_{mask}, \end{aligned} \tag{3}$$

where $\mathcal{L}_{diff}$ and $\mathcal{L}_{mask}$ are the diffusion model training objective and mask loss, respectively; $\mathbf{c}$ is the language guidance features from the referring expressions; $\hat{\mathbf{M}}$ is the mask prediction and $\mathbf{M}$ is the ground truth mask. The parameters λ_1 and λ_2 are loss weights to balance training.

4.2. LGVI-I with MLLM

In the interactive video inpainting task, models are expected to extract valuable information from complex conversations. To overcome this problem, we propose incorporating MLLMs to extend the LGVI from work to LGVI-I (Interactive). MLLMs have demonstrated strong capabilities in visual comprehension and multimodal reasoning, making them well-suited for our proposed interactive video inpainting task. As shown in Fig. 5, the MLLM takes both the image frame and the chat-style user request as inputs, generating the language response and a pair of special indicators: <PROMPT> and </PROMPT>. The hidden states of the last layer between these two indicators are then passed through an MM head, implemented as a two-layer linear layer with activation functions. The transformed features are fed to the cross-attention module to guide the U-Net inpainting process. Mathematically, given the input video $\mathbf{X}$ and user request s, the computation pipeline of the MLLM can be summarized as follows:

$$\begin{aligned} \mathbf{e}_l &= f(s), \quad \mathbf{e}_i = \mathbf{W}_{trans} \cdot g(\mathbf{X}_0), \\ \mathbf{h} &= \mathrm{MLLM}([\mathbf{e}_l, \mathbf{e}_v]), \\ \hat{\mathbf{w}} &= \mathbf{W}_{lm} \cdot \mathbf{h}, \\ \mathbf{h}_p &= \mathbf{W}_{mm} \cdot \mathrm{find\_prompt}(\hat{\mathbf{w}}, \mathbf{h}), \end{aligned} \tag{4}$$

Table 2. **Quantitative results on the referring video inpainting task.** E^*_{warp} denotes $E_{warp}(\times 10^{-2})$.

Method	PSNR ↑	SSIM ↑	VFID ↓	E^*_{warp} ↓
Image Models				
InstructPix2Pix [1]	18.12	0.600	0.361	1.343
Inst-Inpaint [58]	19.00	0.896	0.310	1.206
MagicBrush [66]	20.39	0.725	0.310	0.934
Multi-Stage Video Model				
Inpaint Anything* [61]	22.84	0.728	**0.283**	0.874
One-Stage Video Model				
LGVI (Ours)	**22.85**	**0.756**	0.308	**0.901**

where f is the language token embedding and g is a pre-trained image backbone to extract image features. $\mathbf{W}_{trans}$ is a linear layer that transforms image features into language token space. $\mathbf{h}$ is the last layer's hidden states of the MLLM. $\hat{\mathbf{w}}$ is the predicted language token distribution through the LM head. $\mathbf{W}_{lm}$ is the weights of the LM head. Among the predicted words, we use $\mathrm{find\_prompt}$ function to find the <PROMPT> and </PROMPT> indicator and extract the hidden states that lie between these two indicators. $\mathbf{W}_{lm}$ is the weights of MM head. The MM head transfers the selected tokens into h_p, which is then fed into the U-Net cross-attention module. In this process, h_p serves as vision-aware language guidance for the inpainting process. The training objective of LGVI-I is:

$$\begin{aligned} \mathcal{L}_{diff} &= \mathbb{E}_{\mathbf{X},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[||\epsilon - \epsilon_\theta(\mathbf{v}_t, \mathbf{h}_p, t)||_2^2\right], \\ \mathcal{L}_{lm} &= \mathrm{CrossEntropyLoss}(\hat{\mathbf{w}}, \mathbf{w}), \\ \mathcal{L} &= \lambda_1\mathcal{L}_{diff} + \lambda_2\mathcal{L}_{mask} + \lambda_3\mathcal{L}_{lm}, \end{aligned} \tag{5}$$

where $\mathbf{h}_p$ is the MLLM-enhanced language condition, $\mathcal{L}_{lm}$ is language modeling loss, implemented as the Cross-Entropy Loss, $\mathbf{w}$ is the ground truth sentence, and λ_3 is the loss weight for language modeling loss. By integrating an MLLM into the LGVI framework, the system achieves a higher level of user interactivity. This enables users to guide the visual inpainting process with interactive language instructions, thus establishing a more user-friendly and accessible approach for the interactive video inpainting task.

5. Experiments

5.1. Settings

Datasets and metrics. We use the ROVI dataset test set for both the referring video inpainting and interactive video inpainting tasks. The test set contains 478 videos and 758 objects, each equipped with one referring expression and one interactive expression. During the training of our models, we incorporate a referring image inpainting dataset, GQA-Inpaint [58], to enrich the data vocabulary. We follow video inpainting works [23, 27, 65, 69] to use PSNR and SSIM [49] to assess the statistical similarity between

Table 3. **Results on interactive video inpainting task.** E^*_{warp} denotes $E_{warp}(\times 10^{-2})$. MB represents MagicBrush, and IA* represents Inpaint Anything*. The small numbers on the top 5 rows are compared with the referring video inpainting results. The small numbers on the last row are compared with LGVI.

Method	PSNR ↑	SSIM ↑	VFID ↓	E^*_{warp} ↓
Image Models				
InstructPix2Pix [1]	16.53(-1.59)	0.558(-0.042)	0.391(-0.003)	1.789(-0.446)
Inst-Inpaint [58]	18.96(-0.04)	0.702	0.314(-0.004)	1.047
MagicBrush [66]	20.46	0.728	0.311(-0.001)	0.901
Multi-Stage Video Model				
IA* [61]	20.64(-2.20)	0.664(-0.064)	0.312(-0.029)	1.182(-0.308)
One-Stage Video Model				
LGVI (Ours)	20.70(-2.15)	0.707(-0.049)	0.332(-0.024)	1.191(-0.290)
MLLM-Enhanced Two-Stage Model				
MB + MLLM	20.37	0.726	0.313	1.004
IA* + MLLM	21.37	0.722	0.291	0.875
LGVI + MLLM	21.45	**0.738**	0.311	0.923
MLLM-Enhanced End-to-End Model				
LGVI-I (Ours)	**22.24**(+1.54)	0.732(+0.025)	**0.299**(+0.033)	**0.867**(+0.324)

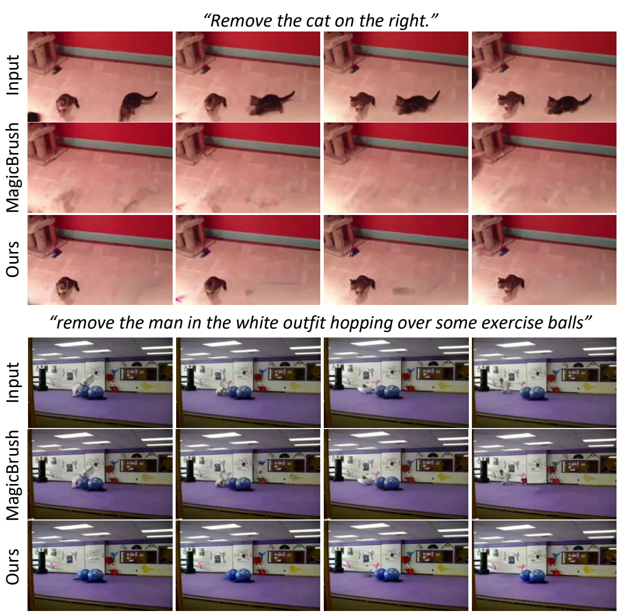


Figure 6. Qualitative comparison between LGVI and MagicBrush [66] on the referring video inpainting task.

predicted results and ground truth. Additionally, we use VIFD [48] to measure the perceptual similarities. To assess the temporal consistency and smoothness of the generated videos, we also apply the E_{warp} metric [14].

Baselines. For the baselines, we select three language-driven image editing methods: InstructPix2Pix [1], Inst-Inpaint [58], and MagicBrush [66]. It is worth noting that InstructPix2Pix and MagicBrush are pre-trained on extensive image editing datasets. Inst-Inpaint is proposed to perform referring image inpainting on images. We also compare with Inpaint Anything, a multi-stage method for one-click video inpainting. It uses Segment Anything [13] and

Figure 7. Examples of referring to multiple objects in one sentence and referring to nonexistence objects.

OSTrack [57] to produce segmentation masks based on user click, followed by inpainting the masked area using inpainting models [64]. We implement Inpaint Anything*, which facilities Inpaint Anything [61] with GroundingDINO [28], enabling it to process language inputs.

Implementation details. We initialize the U-Net weights from MagicBrush [66]. The newly introduced modules are trained from scratch. During training, we sample video and image data at a ratio of 3 : 1. For the MLLM, we adopt LLaVA-7B [26]. The learning rates are 3e-5, 1e-4, and 1e-4 for U-Net, mask decoder, and tuned parameters in MLLM, respectively. The loss weights are set to $\lambda_1 = 1$, $\lambda_2 = 0.001$, $\lambda_3 = 0.1$. These weights differ significantly due to the different types of losses they represent. The input and output video sizes are set to 512×320, and the video length is 24. For LGVI, we train 50 epochs on the ROVI dataset with a batch size of 32 for videos and 768 for images. For LGVI-I, we load the LGVI checkpoint and fine-tune it for 50 epochs under the same batch size. All experiments are carried out on 8 NVIDIA A100 GPUs.

5.2. Referring Video Inpainting

Quatitative results. We report quantitative results on the referring video inpainting task. Compared with baseline models, our model is the first one-stage language-driven video inpainting model. As shown in Tab. 2, our model outperforms MagicBrush [66] in all metrics and achieves on-par results with Inpaint Anything* [61], even if Inpaint Anything* uses a mask-based inpainting model [64]. The results demonstrate the effectiveness of our model.

Qualitative results. Figure 6 shows qualitative results. We compare with MagicBrush [66], a robust generalized language-driven image editing model. In the first example, where the language refers to the cat on the right, the MagicBrush model removes all the cats in the scene, while our model successfully inpaints the right cat. In the second example, the referring expression becomes more complex. MagicBrush struggles to identify the object requiring inpainting and removes the wrong object (the balls) in

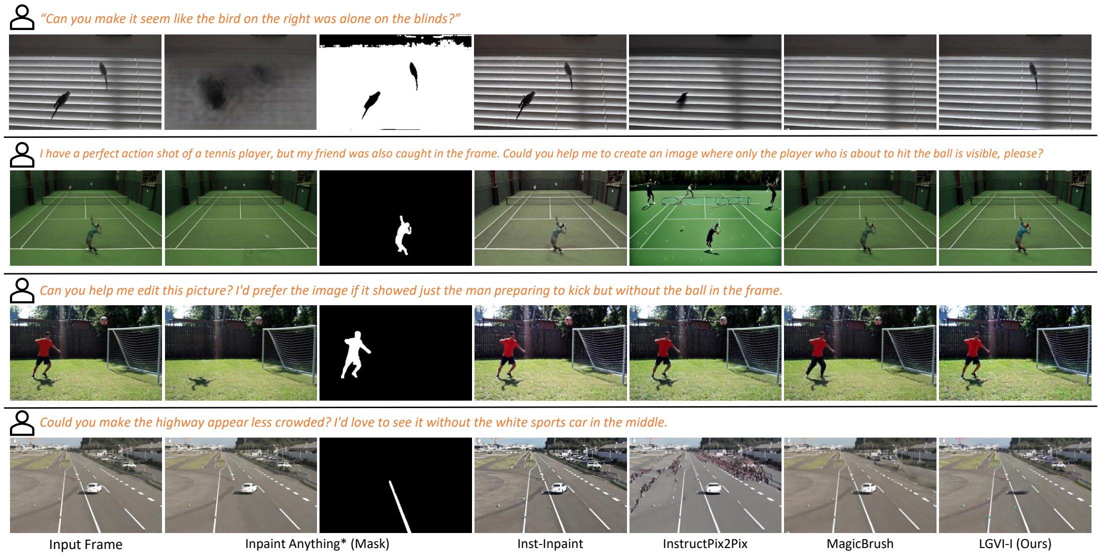


Figure 8. **Qualitative comparison between LGVI-I and baseline models on the interactive video inpainting task**. The chat-style conversation inputs are listed above each row. Columns 2 and 3 are the results and predicted masks from Inpaint Anything*. It removes the inaccurate objects according to the wrongly predicted masks due to the difficult interactive language inputs.

the last frame. In contrast, our model generates a plausible output, demonstrating its superior performance in handling complex language-driven inpainting tasks. Furthermore, in Fig. 7, we compare with Inpaint Anything* on sentences referring to multiple objects or nonexistent objects. Inpaint Anything is driven by a simple combination of referring segmentation and video inpainting models. Thus, it is fixed to produce one mask for each sentence. When referring to multiple or nonexistent objects, it outputs inaccurate results, while our model produces the correct output. This demonstrates the robustness of the language-driven setting.

5.3. Interactive Video Inpainting

Quatitative results. We report the interactive video inpainting task results in Tab. 3. As shown in the top 5 rows, when the models are trained using referring expressions, their performance drops correspondingly in this task. This is intuitive because interactive expressions are much longer and more implicit. For the MLLM-Enhanced Two-Stage Models, we combine the baseline models with an MLLM in a zero-shot manner. The interactive inputs are transferred into shorter referring expressions by simply prompting the MLLM. These models exhibit improved performances. Our LGVI-I model achieves the highest performance, demonstrating the effectiveness of the proposed architecture.

Qualitative results. Figure 8 presents examples of the interactive video inpainting task. The user requests pose a significant challenge and complexity for the baseline models to comprehend. In particular, Inpaint Anything* predicts incorrect masks, leading to inaccurate results. Similarly, other diffusion-based models struggle to interpret the users' intentions accurately, resulting in less satisfactory outcomes. In contrast, our LGVI-I model, which harnesses the power of MLLM, consistently delivers the best performance in these challenging scenarios. This underscores the superiority of our proposed approach. More detailed ablations can be seen in the supplementary due to the page limitation.

6. Conclusion

In this paper, we propose a novel language-driven video inpainting task that uses language to guide inpainting areas. For training and evaluation, we collect a video dataset, namely ROVI. Comprehensive statistics demonstrate the uniqueness and diversity of our dataset, especially the chat-style interactive conversations generated by powerful LLMs and MLLMs. We further propose a diffusion-based baseline model, LGVI. Quantitative and qualitative experimental results show the effectiveness and robustness of our model. We hope our proposed benchmark and baselines can provide valuable insights into multi-modal models of low-level vision. In addition, there are several challenges to solve, including scalability and generalization of the model. We list the more discussion in the appendix.

Acknowledgement. This project is supported by the National Key Research and Development Program of China (No.2023YFC3807600) and the National Key R&D Program of China (No.2022ZD0161600). This study is also partially supported by the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 6, 7

[2] 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. *NeurIPS*, 2020. 3

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 2, 3

[4] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *ICCV*, 2019. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[6] Mounira Ebdelli, Olivier Le Meur, and Christine Guillemot. Video inpainting with short-term windows: application to object removal and error concealment. *TIP*, 2015. 1

[7] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018. 3, 4

[8] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 3

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[10] Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G Schwing. Proposal-based video completion. In *ECCV*, 2020. 1, 2

[11] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*, 2019. 3

[12] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 7

[14] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 7

[15] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *ICCV*, 2019. 1

[16] Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. Short-term and long-term context aggregation network for video inpainting. In *ECCV*, 2020. 2

[17] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 3

[18] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *CVPR*, 2020. 2

[19] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-aware transformer for large hole image inpainting. In *CVPR*, 2022. 2

[20] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *ECCV*, 2020. 2

[21] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Guangliang Cheng, Pang Jiangmiao, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv pre-print*, 2023.

[22] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *CVPR*, 2024. 2

[23] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *CVPR*, 2022. 1, 2, 4, 6

[24] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 2

[25] Guilin Liu, Aysegul Dundar, Kevin J Shih, Ting-Chun Wang, Fitsum A Reda, Karan Sapra, Zhiding Yu, Xiaodong Yang, Andrew Tao, and Bryan Catanzaro. Partial convolution for padding, inpainting, and image synthesis. *TPAMI*, 2022. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3, 7

[27] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *ICCV*, 2021. 2, 6

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 7

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 3

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[33] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive vision transformers for efficient image recognition. In *CVPR*, 2022. 2

[34] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2

[35] OpenAI. Gpt-4 technical report, 2023. 3

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[37] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 3

[38] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*, 2023. 3

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[40] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 5

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2

[43] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. 3, 4

[44] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 5

[45] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. In *CVPR*, 2023. 2

[46] 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. 3

[47] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *AAAI*, 2019. 1, 2

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *NeurIPS*, 2018. 7

[49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[50] Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation. *ICCV*, 2023. 2

[51] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *TIP*, 2024.

[52] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, et al. Towards open vocabulary learning: A survey. *T-PAMI*, 2024. 2

[53] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 2, 5

[54] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023. 3

[55] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 3

[56] Shilin Xu, Haobo Yuan, Qingyu Shi, Lu Qi, Jingbo Wang, Yibo Yang, Yining Li, Kai Chen, Yunhai Tong, Bernard Ghanem, Xiangtai Li, and Ming-Hsuan Yang. Rap-sam: Towards real-time all-purpose segment anything. *arXiv preprint*, 2024. 2

[57] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 1, 7

[58] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 3, 6, 7

[59] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *CVPR*, 2022. 2

[60] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 2

[61] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 1, 6, 7

[62] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *arXiv preprint*, 2024. 2

[63] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*, 2023. 2, 3

[64] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020. 1, 7

[65] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *ECCV*, 2022. 2, 6

[66] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023. 2, 6, 7

[67] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 3

[68] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660*, 2023. 2

[69] Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *ICCV*, 2023. 1, 2, 6

[70] 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. 2, 3