Text-Guided Video Amodal Completion

Minh Tran[1] Winston Bounsavy[1] Taisei Hanyu[1] Thang Pham[1]
Khoa Vo[1] Tri Nguyen[2] Ngan Le[1]
[1]University of Arkansas, [2]Coupang, Inc.

https://uark-aicv.github.io/TextGuidedVAC

Abstract

Amodal perception enables humans to perceive entire objects even when parts are occluded, a remarkable cognitive skill that artificial intelligence struggles to replicate. While substantial advancements have been made in image amodal completion, video amodal completion remains underexplored despite its high potential for real-world applications in video editing and analysis. In response, we propose a video amodal completion framework to explore this potential direction. Our contributions include (i) a large-scale synthetic dataset for video amodal completion with text description for the object of interest. The dataset captures a variety of object types, textures, motions, and scenarios to support zero-shot transferring on natural videos. (ii) A diffusion-based text-guided video amodal completion framework enhanced with a motion continuity module to ensure temporal consistency across frames. (iii) Zero-shot inference for long video, inspired by temporal diffusion techniques to effectively manage long video sequences while improving inference accuracy and maintaining coherent amodal completions. Experimental results shows the efficacy of our approach in handling video amodal completion, opening potential capabilities for advanced video editing and analysis with amodal completion.

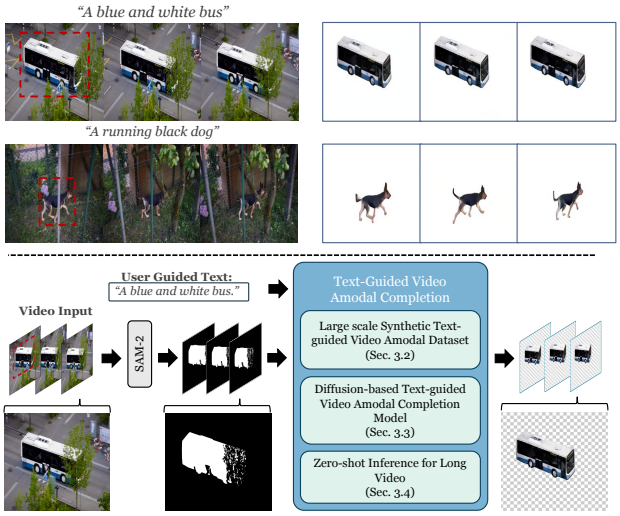

Figure 1. An overview of our text-guided video amodal completion pipeline. Given an input video, users select an object of interest in the first frame and provide a text description of the expected output. Our pipeline then generates a completed video, filling in the missing shape and texture of the object. (Top) Zero-shot transfer results on natural videos using our method. (Bottom) A summary of our proposed pipeline and three key contributions.

1. Introduction

Humans possess an extraordinary ability to perceive objects as complete entities, even when parts are obscured. In everyday life, objects frequently block one another from view. Yet, we effortlessly identify and reconstruct their hidden portions—a capability known as amodal perception or amodal completion [26, 47]. Human's visual system achieves this by relying on shape continuity, symmetry [47], and a deep familiarity with the world around us [55]. Replicating this cognitive process of amodal completion presents a significant challenge for artificial intelligence (AI), despite recent advances in computer vision. Achieving amodal completion in AI could benefit diverse applications, including robotics [1, 36], autonomous driving [35], and augmented reality [33]. Similar to other human visual abilities, amodal perception has inspired the development of AI algorithms designed to mimic this capability. Research has initially focused on amodal segmentation [12, 13, 15, 30, 46, 54], where models attempt to obtain the object's complete shape. More recently, the advent of denoising diffusion models [19, 40] has spurred progress in amodal content completion [33, 52, 56].

Despite significant advancements in image amodal completion [33, 52, 56], research in video amodal completion remain unexplored. This is primarily due to the challenges posed by temporal dimension and dynamic occlusions in video data, despite the broad applications in fields like robotics [1], autonomous driving [35], video editing [31],

and content creation [5]. Unlike static images, video requires models to simultaneously track and complete occluded objects consistently across frames, enabling seamless continuity and realistic rendering of occluded regions. Additionally, a key limitation of prior image amodal completion datasets is their lack of auxiliary information to describe occluded content [11]. This absence is particularly problematic in cases of significant occlusion, where inferring and completing hidden parts becomes ambiguous or ill-posed. To address these challenges, we present three key contributions in this work as follows:

i. Large-scale Synthetic Text-guided Video Amodal Completion Dataset: We introduce a large-scale synthetic video amodal completion dataset. This dataset encompasses diverse object categories and scenarios, providing knowledge on various shapes, textures, and motions to facilitate zero-shot transferring. To address the inherent ambiguity in amodal completion, we enhance the dataset with detailed textual descriptions of occluded regions, offering explicit guidance for accurately completing hidden content.

ii. Diffusion-based Text-guided Video Amodal Completion Model: We propose a novel framework for video amodal completion that generates complete object shapes, textures, and motions. As illustrated in Figure 1, given an input video, our proposed text-guided video amodal completion aims to extract the object and fill in occluded areas with semantically coherent information. To the best of our knowledge, this is the first exploration of amodal completion in videos. Inspired by the recent advancements video generation [3, 20, 23], we leverage diffusion models to establish our baseline. Our approach employs a two-phase training strategy: frame-level training and motion training. In the first phase, we train a denoising UNet at the frame level to effectively capture spatial features. In the second phase, we focus on training motion layers while keeping the frame-level layers frozen, ensuring temporal coherence across frames.

iii. Zero-shot Inference for Long Video: Inspired by MultiDiffusion [2], Temporal Diffusion [59], our approach manages long videos by dividing them into training-sized clips. The resulting completions are then integrated to ensure consistency across the entire video.

Experimental results show that our framework effectively outperforms the existing frame-level amodal completion methods as well as show the capability of zero shot-transferring to natural videos, unlocking new possibilities for advanced video editing and analysis through video amodal completion.

2. Related Work

2.1. Amodal Completion and Segmentation

Amodal visual understanding has been explored in various aspects to complement normal visual understanding, often resulting in outputs obscured by foreground objects. For example, amodal segmentation generates a complete mask of a particular object [25, 31, 35, 39, 60]. Amodal detection predicts entire objects, including hidden parts [22, 24]. More recently, amodal completion aims to generate the complete shape of an object [8, 33, 57]. The first two tasks have been well-explored, mainly due to advancements in model domain methods for visible mask problems. Furthermore, thanks to numerous closed-world datasets [24, 25, 31, 35, 57, 60] and large synthetic datasets [8], amodal segmentation methods have developed significantly, such as PCNet [57] or AISFormer [46]. Controlling and generating whole objects is a more challenging task due to the non-trivial nature of conditioning on visible masks for generation. Pix2gestalt [33] addresses this challenge preliminarily by fine-tuning a large-scale diffusion model on a synthetic dataset. Despite significant advancements in image amodal completion [33, 52, 56], there has been limited exploration of video amodal completion. In this work, we aim to bridge this gap by studying the challenges and opportunities within the video amodal completion problem.

2.2. Diffusion Models

Denoising Diffusion Probabilistic Model [19], or DDPM, has emerged as one of the most powerful methods among generative models. It is widely applied in various computer vision tasks, including image [7, 18, 45] and video generation [3, 17, 20], motion generation [23, 59], 3D generation [6, 32, 51], and out of computer vision like waveform generation [28]. The breakthrough starts with [7] demonstrating that diffusion models can outperform GANs [16] in image synthesis. This is followed by the success of Stable Diffusion [40], trained on LAION-5B [43], which offers improved computational efficiency. One of the key features of Stable Diffusion is that we can guide it using text prompt, which leads to many advances in image editing [4, 14, 42]. Recently, diffusion models also have been explored in video syntheis. Video Diffusion Models (VDM) [21] naturally extend the concept of text-to-image diffusion models by training on both image and video datasets. Imagen Video [20] creates a cascade of video diffusion models and incorporates spatial and temporal super-resolution techniques to produce high-resolution, time-consistent videos. Make-A-Video [44] builds on text-to-image synthesis models and employs unsupervised video data to enhance performance. Gen-1 [9] expands on SD by introducing a structure- and content-guided video editing approach, using either visual or textual descriptions of the desired outcome. Tune-A-

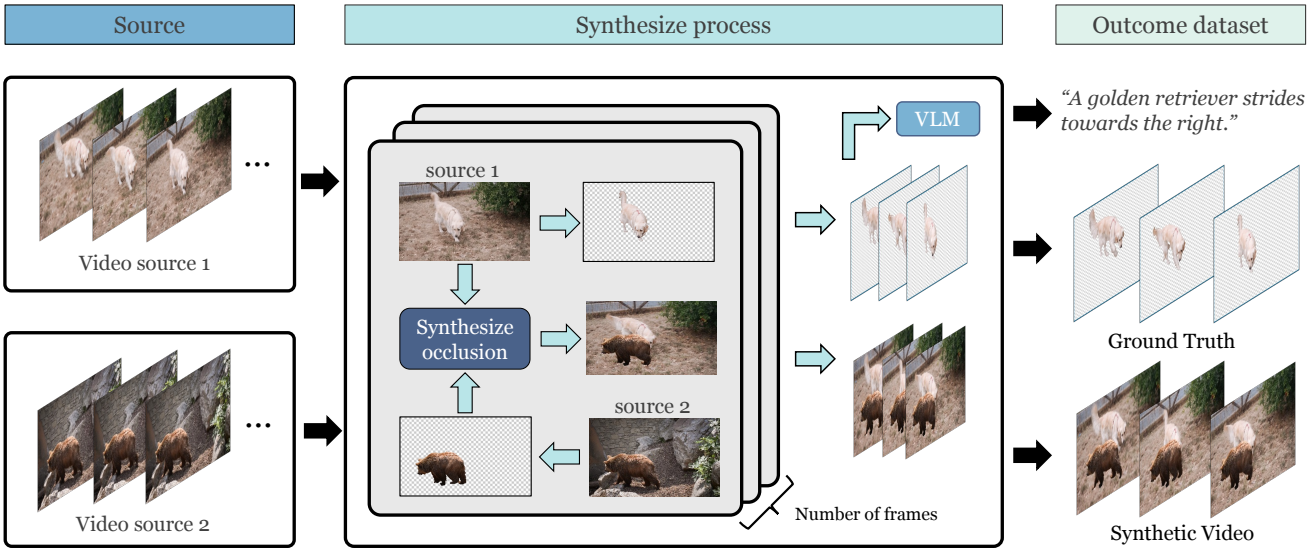

Figure 2. **Synthesizing training data process.** We selected videos with unoccluded objects and used provided masks to isolate object pixels. These were used to synthesize occlusion videos and create corresponding amodal completion ground truth. A vision language model (e.g. BLIP [29] is utilized to generate ground truth's text description.

Video [50] introduces a novel task of one-shot video generation, extending SD with a single reference video.

3. Video Amodal Completion

Given an RGB video $\mathbf{v} = \{\mathbf{x}_i\}_{i=1}^{N}$ and an initial prompt $\mathbf{p}_0$ (which can be a text description, mask, or bounding box) specifying an object of interest $\mathbf{o}$ in the first frame $\mathbf{x}_0$, the amodal video completion task involves identifying both the visible and occluded parts of this object $\mathbf{o}$ throughout the video sequence $\mathbf{v}$. Formally, let $\mathbf{v}_{\text{out}}$ denote the output video depicting the complete object of interest $\mathbf{o}$, we have:

$$\mathbf{v}_{\text{out}} = f_\theta(\mathbf{v}, \mathbf{p}_0) \tag{1}$$

Here, $f_\theta(\cdot)$ denotes an estimator function, such as a conditional diffusion model. The goal is for the visible portions of the object $\mathbf{o}$ in $\mathbf{v}_{\text{out}}$ to accurately match the corresponding visible mask of $\mathbf{o}$ in the input video $\mathbf{v}$. Additionally, the completed (occluded) portions should integrate seamlessly, maintaining contextual consistency and avoiding any physically implausible object configurations.

3.1. Preliminaries

Diffusion models [19] aim to learn the data distribution $p(\mathbf{x})$ by sequentially denoising images. In the forward process, noise is gradually added to an image $\mathbf{x}$ over T time steps, transforming it into a sample with nearly Gaussian noise. In the reverse process, the model learns to remove this noise over T steps. At each step $t = [1, T]$, a neural network predicts the noise $\epsilon_\theta(\mathbf{x}^t, t)$ for the noisy image $\mathbf{x}^t$. Unlike standard diffusion models that operate directly on image pixels, latent diffusion models (LDMs) [40] work in the latent space of pre-trained autoencoders. Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, an encoder E encodes $\mathbf{x}$ into a latent representation $\mathbf{z} = E(\mathbf{x})$, and a decoder D reconstructs $\mathbf{x}$ from $\mathbf{z}$ as $\hat{I} = D(\mathbf{z})$. In this framework, the autoencoder functions as a time-conditional UNet [41], denoted as $\epsilon_\theta(\mathbf{z}_t, t)$, where t is a specific time step and $\mathbf{z}_t$ is the latent representation at that step. To incorporate an input condition y, such as images, masks, text, LDMs integrate cross-attention layers [48] into the denoising UNet, enabling y to map to the intermediate layers of the UNet [40].

3.2. Video Amodal Completion Dataset

A significant obstacle in amodal completion research is the scarcity of natural image datasets that include ground truth for amodal scenarios. Previous studies on image amodal completion [33, 52, 57] address this by generating pseudo-occluded images. However, amodal completion dataset for video-level have been unexplored. To tackle the challenge, we develop a large-scale synthetic dataset tailored for this specific purpose. We aim to ensure our dataset encompassed a wide range of common objects and scenarios, enabling the extraction of knowledge related to various shapes, textures, and motions to support zero-shot transferring ability. Previous amodal completion studies [33, 52, 57] typically create pairs of images where a single, whole (unoccluded) object is overlaid with an occluder, producing a pseudo-occluded image as input and its corresponding whole counterpart as ground truth. We adopted and expanded upon these strategies to create a compre-

Table 1. High level statistic of our synthetic dataset regarding each source.

	DAVIS [34]	**YTVOS** [53]	**LaSOT** [10]
Mean #frames	57	113	1,797
# object classes	9	65	60
# instances	24	332	1,044
# synthesized video	120	1,600	5,220

hensive video amodal completion dataset, resulting pairs of pseudo-occluded videos as inputs and their corresponding videos of single whole objects as ground truth. However, a notable limitation of prior image amodal completion datasets is the lack of auxiliary information to describe the occluded object, particularly in cases of significant occlusion, making it challenging to infer the hidden parts[52]. To address this, we enhanced our dataset by including a textual description for each video data point, which describe the single whole object in the ground truth and serves as an extra conditional input. In specific, we sourced videos from video object segmentation datasets such as DAVIS [34], YTVOS [53], and LaSOT [10]. These datasets were chosen because their videos typically feature a single object of interest with minimal occlusion and significant motion variation, making them suitable as ground truth for amodal tasks. For DAVIS [34] and YTVOS [53], we first selected videos featuring objects without occlusion and used the provided masks throughout the video frames to isolate the object pixels. These were then used to create synthesized occlusion videos and corresponding amodal completion ground truth. As illustrated in Figure 2, given videos of a dog walking and a camel walking, we utilized their annotated masks to create occlusion scenarios by overlapping them, generating one occlusion video as input and retaining the complete, unoccluded pixels as ground truth. The LaSOT [10] dataset, originally a video object tracking dataset lacking annotated masks, was incorporated due to its extensive variety in object types, motions, and large video count, with most videos containing a single object. We leveraged SAM [27] to obtain initial segmentation masks using the annotated bounding boxes and subsequently refined these masks to correct any inaccuracies. For the textual description, we leverage BLIP-2 [29] to generate the textual caption given the first frame of the ground truth video.

Tab. 1 provides detailed statistics of the dataset, breaking down the object classes, number of instances, and total synthesized videos for each source.

3.3. Text-guided Video Amodal Completion

Given the problem definition in 1, with an RGB video $\mathbf{v} = \{\mathbf{x}_i\}_{i=1}^N$ and an initial prompt $\mathbf{p}_0$ (e.g. points, bounding box, or mask) specifying an object of interest $\mathbf{o}$ in the first frame $\mathbf{x}_0$, we utilize SAM-2 [38] to obtain the visible masks $\mathbf{v}_m = \{\mathbf{m}_i\}_{i=1}^N$ of $\mathbf{o}$ across all frames.

A key limitation of existing image amodal completion datasets is their lack of auxiliary information to describe occluded objects, makes inferring the hidden parts challenging especially in cases of significant occlusion [52]. To alleviate this, we propose incorporating a text prompt y into the video amodal completion task to describe the desired whole object. This approach leverages the capability of pretrained T2I diffusion models, enabling them to generate images with textual prompts for improved amodal completion.

To the best of our knowledge, there are no existing works that have explored amodal completion in video data. In this study, we propose a baseline method for this task. Inspired by recent advancements in diffusion models for video generation [3, 20, 23], we utilize a diffusion model to establish our baseline. Our approach follows a common two-phase training strategy: first, training a denoising UNet at the frame level to capture spatial features, and then incorporating motion training to ensure temporal coherence. Figure 3 illustrates the overall training scheme of our method.

Frame-level Training. Inspired by [33], we fine-tune a conditional diffusion model (e.g., Stable Diffusion [40]) to perform frame-level amodal completion. Specifically, we optimize the following latent diffusion objective:

$$\min_{\theta_k} \mathbb{E}_{\mathbf{z},\epsilon,t}[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathcal{E}(\mathbf{x}_i), t, \mathcal{E}(\mathbf{m}_i), \mathcal{C}(y))\|_2^2] \quad (2)$$

where ϵ is a denoiser U-Net, θ_k represents the frame-level parameters, and $\epsilon \sim \mathcal{N}(0,1)$ denotes Gaussian noise. The variable $\mathbf{z} \sim \mathcal{E}(\mathbf{x}_i)$ is the VAE embedding of the video frame $\mathbf{x}_i$, and $t \in [0, 1000]$ is the diffusion time step. $\mathbf{z}_t$ is the noised embedding of the amodal target object $\mathbf{o}$ at the ith frame. The term $\mathcal{C}(y)$ represents the CLIP text embedding of the input text prompt y, while $\mathcal{E}(\cdot)$ is the VAE encoder of the diffusion model.

Two guiding information flows contribute to our objective. The first is conditional information provided through a cross-attention mechanism $\mathcal{C}(y)$, which steers the model towards the desired features of the complete object specified by the text prompt y. The second flow involves concatenating $\mathbf{z}_t$, and $\mathcal{E}(\mathbf{m}_i)$ to enforce adherence to the visible parts of the object of interest, ensuring the model remains consistent with the given data.

Motion Training. After pretraining at the frame level, we focus on modeling the motion dynamics between frames. This step aims to create smooth motion generation and enhance completion quality, allowing frames with less occlusion to share features with those that have more occlusion. To leverage the knowledge from frame-level training, it is advantageous to inflate the network so that image layers can handle video frames independently [23]. Following recent approaches [3, 20, 23], we modify the model to accept video tensors $\mathbf{v} \in \mathbb{R}^{B\times N\times C\times H\times W}$, where B repre-

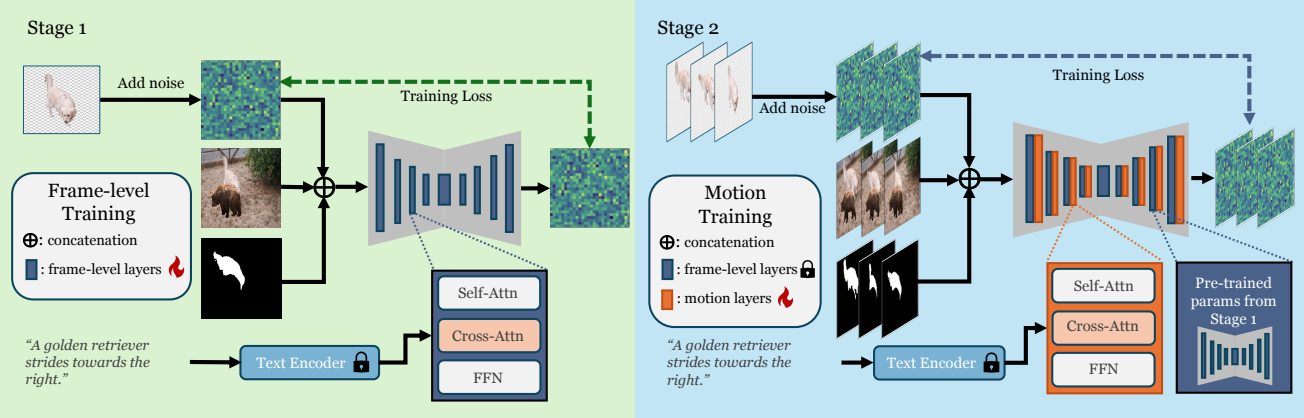

Figure 3. **Training pipeline of the proposed method.** Our approach follows a common two-stage training strategy: first, training a denoising UNet at the frame level to capture spatial features, and then incorporating motion training to ensure temporal coherence.

sents the batch axis and N the time axis. Internally, when feature maps pass through image layers, the temporal axis N is reshaped into the batch axis B, allowing the network to process each frame independently. After processing, the feature map is reshaped back into a 5D tensor.

For motion modeling, we introduce a motion module that processes the temporal dynamics by reshaping the spatial axes h and w into the batch axis and reverting them after processing. Inspired by [23], we design this module using a Transformer architecture [48]. After the image layers, feature maps from video frames $\{\mathbf{z}_i\}_{i=1}^{N} \in \mathbb{R}^{(b\times h'\times w')\times c'}$, where the spatial dimensions are merged into the batch axis, are obtained. We employ relative position embeddings to maintain the order of frames, enabling the temporal attention block to capture temporal coherence. During motion training, we freeze the image layers and train the motion module by optimizing the following objective:

$$\min_{\theta_m} \mathbb{E}_{\mathbf{z},\epsilon,t}[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathcal{E}(\mathbf{v}), t, \mathcal{E}(\mathbf{v}_m), \mathcal{C}(y))\|_2^2] \quad (3)$$

where θ_m represents the parameters of the motion module.

3.4. Zero-shot Inference for Long Video

While the above pipeline can theoretically handle videos of arbitrary length N, the model may experience significant quality degradation when generating videos longer than those used in training [23, 59]. Inspired by approaches such as MultiDiffusion [2], which generates high-resolution images seamlessly composed of multiple patches, and Temporal Diffusion [59], which handles video inpainting by generating multiple clips, we adapt Temporal Diffusion [59] for our video amodal completion task. In specific, we divide the video into smaller training-sized clips N', denoted as $\mathbf{v}^i$ for $i \in [1, N']$, using a stride s. At each denoising timestep t, our model is applied N' times to produce N' output clips, represented as $\mathbf{v}^i_{out_t} = \epsilon_\theta(\mathbf{z}_{t-1}, \mathcal{E}(\mathbf{v}^i), t-1, \mathcal{E}(\mathbf{v}^i_m), \mathcal{C}(y))$. Overlapping frames between these clips are averaged based on the number of times they are processed, ensuring consistency across the entire video sequence.

4. Experimental Result

4.1. Implementation Details

Training. Our model is built upon Stable Diffusion [40] version 1.5 from the Diffusers library [49]. Initially, we train the frame-level layers of our U-Net using the Pix2Gestalt dataset [33], which consists of approximately 800,000 data samples. This stage involves 500,000 training steps with a batch size of 16, where the input image resolution is 256×256. We employ DDIM [45] sampling with denoising steps $T = 1000$. Following frame-level training, we freeze the frame-level layers and proceed to train the motion layers using our synthesized dataset described in Sec. 3.2. The dataset is split 80-20 for training and validation. We conduct 500,000 training steps with a batch size of 16 and train on sequences of 14 frames per sample, each with a resolution of 256×256. Both training stages use a learning rate of 0.0005. Our experiments are performed on eight NVIDIA RTX A6000 GPUs, each with 48 GB of memory.

Inference. During inference, we conduct zero-shot testing on arbitrary video sequences using Temporal Diffusion (Sec. 3.4) with a stride of $o = 4$ and clip length of 14 frames. We utilize DDIM [45] with 30 denoising steps and apply a classifier guidance scale of 7.5.

4.2. Comparison with related methods

Prior works. To the best of our knowledge, no existing work has explored amodal completion at the video level. In this study, we benchmark our method against frame-level amodal completion techniques to highlight its ability to maintain temporal consistency while ensuring high com-

Table 2. **Quantitative results.** We compare our method against frame-level amodal completion methods Pix2gestalt [33], ProgressiveAmodal [52], and evaluate generated results using different metrics, including CLIP [37] (high-level), LPIPS [58] (low-level), and Temporal Consistency [9].

Method	Easy Cases				Hard Cases			
	CLIP↑	LSIPS↓	TC↑	User Preference	CLIP↑	LSIPS↓	TC↑	User Preference
ProgressiveAmodal [52]	0.88	0.14	0.92	0.10	0.90	0.18	0.85	0.11
Pix2Gestalt [33]	0.89	0.12	0.92	0.12	0.90	0.17	0.87	0.08
Ours	**0.93**	**0.06**	**0.96**	**0.78**	**0.92**	**0.14**	**0.93**	**0.82**

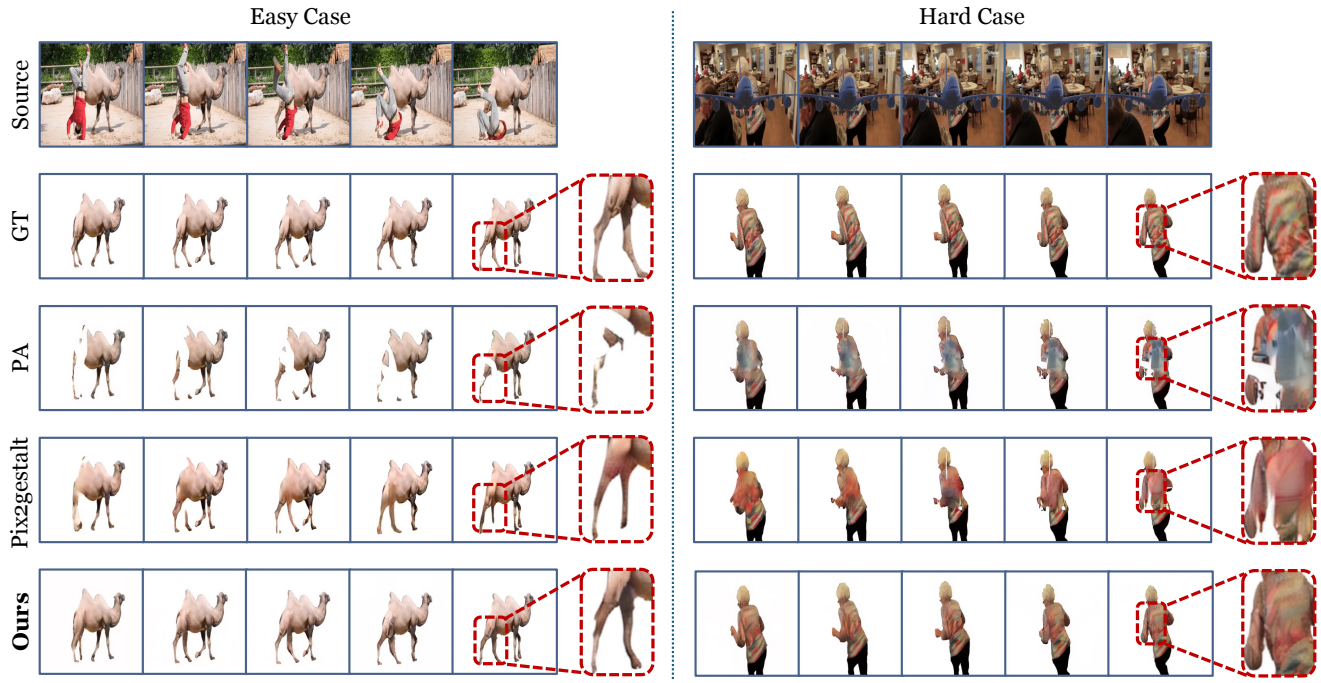

Figure 4. **Qualitative comparison.** We compare our method against frame-level amodal completion methods Pix2gestalt [33], ProgressiveAmodal (PA) [52]. Our proposed approach successfully complete the object with realistic context while retaining the visible parts of objects. Moreover, our results demonstrate higher perceptual completeness and temporal consistency, outperforming other methods in our comparison.

pletion quality. We compare it with recent state-of-the-art (SOTA) frame-level amodal completion methods, specifically Pix2gestalt [33] and ProgressiveAmodal [52].

Validation Set. Our proposed synthetic dataset (Sec. 3.2) is divided into an 80-20 split for training and validation. The validation set comprises 1,400 videos for quantitative evaluation. Within this set, we classify video frames into easy cases (less than 50% occlusion) and hard cases (greater than 50% occlusion), following the approach outlined in [52].

Metrics. Following [52], we employ CLIP [37] for assessing high-level image similarity and LPIPS [58] for evaluating low-level image similarity. Additionally, we measure Temporal Consistency (TC) by calculating the cosine similarity between consecutive frames within the CLIP-Image feature space, as done in [9, 59]. To further measure perceptual video quality, we conduct a user preference study involving 20 university students. Participants are shown 20 input videos (10 easy cases, 10 hard cases) alongside generated amodal completion videos from each method, and they are asked to vote for the video that appears most complete and realistic. Detailed study protocols and user demographics are provided in the Supplementary Materials.

Quantitative comparison. Tab. 2 demonstrate that in both 'Easy' and 'Hard' scenarios, our method consistently surpasses Pix2Gestalt and ProgressiveAmodal. When evaluated using the high-level image similarity metric CLIP, our approach outperforms both baselines, achieving scores of 0.93 compared to 0.89 in easy cases and 0.92 compared to 0.90 in hard cases. Although these improvements appear modest, this is expected in the task of amodal com-

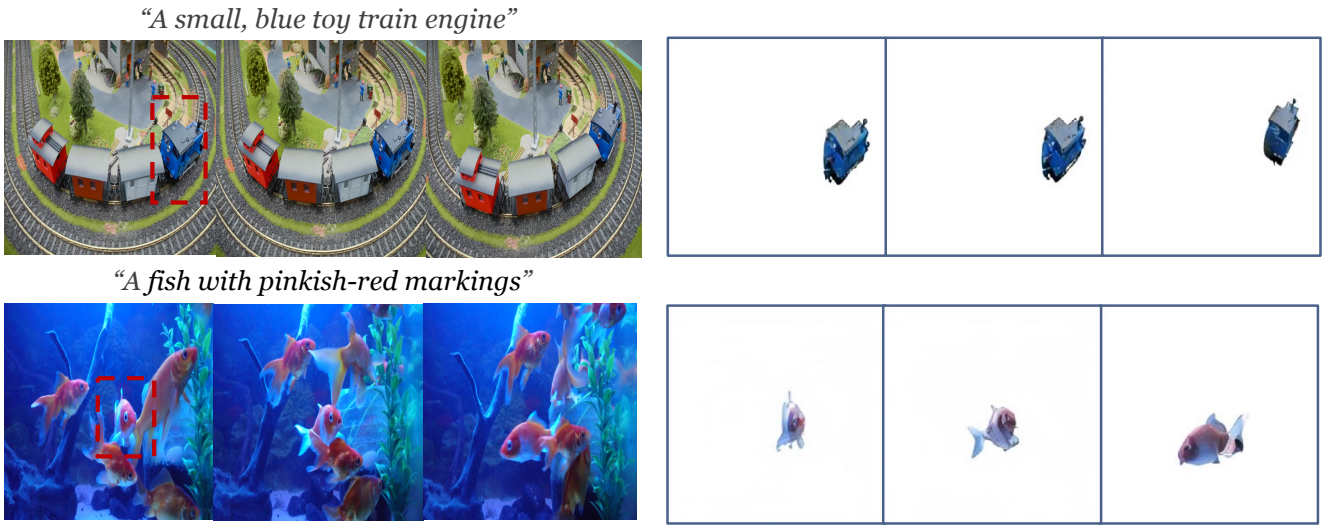

Figure 5. **Qualitative results showing zero-shot amodal completion on natural videos.** The top row depicts a toy train partially hidden by surrounding objects (left). Given a query "A small, blue toy train enginge", the model reconstructs the missing parts of the blue toy train and renders isolated views of the completed shape (right). Similarly, the bottom row shows a fish with pinkish-red markings as our target, where parts of its body are occluded by another fish; the model takes the query "A fish with pinkish-red markings" and then completes the fish's full shape.

pletion because even suboptimal generated parts often align well with the visible sections' colors and shapes, resulting in high CLIP similarity scores. In contrast, significant advancements are evident in the low-level image similarity metric LPIPS, where our method shows clear superiority: 0.06 versus 0.12 in the 'Easy' case and 0.14 versus 0.17 in the 'Hard' case. This improvement is attributed to the training with the motion module, which enhances coherence between frames and minimizes abrupt changes during occlusion. This advantage is further highlighted by the Temporal Consistency metric, where our method also outperforms the other two, scoring 0.96 versus 0.92 in the 'Easy' case and 0.93 versus 0.87 in the 'Hard' case. In the User Preference study, our method achieves significantly higher preference scores in both "Easy Cases" (0.78) and "Hard Cases" (0.82), substantially outperforming both Pix2Gestalt (0.10 and 0.11) and ProgressiveAmodal (0.12 and 0.08). These results quantitatively demonstrate that users overwhelmingly preferred our approach for its superior completeness, realism, and perceptual quality in both straightforward and challenging scenarios.

Qualitative comparison. Visual comparisons in Figure 4 show our method's superior performance against existing approaches. While prior works, such as Pix2Gestalt [33], can produce reasonable high-level completions, they fall short in generating fine details accurately. In contrast, our model demonstrates higher accuracy, particularly in reconstructing occluded details. For example, our approach more precisely reconstructs challenging areas, such as the camel's legs and the back of the elderly woman.

4.3. Zero-shot Inference on Natural Videos.

In Figure 5, we present qualitative results demonstrating zero-shot performance on natural videos. As shown, our model effectively completes amodal occlusions in real-world video sequences. This highlights the value of our synthetic amodal completion dataset, which captures diverse object types, textures, motions, and scenarios, enabling robust zero-shot transfer to natural video contexts. Additional examples of zero-shot inference on natural videos are provided in the supplementary materials.

4.4. Ablation Study

Effect of Condition Streams. Our network incorporates two distinct streams for conditioning. The first stream involves concatenating the visible mask feature with noise and the input image feature. The second stream uses encoded text features to describe the desired output. As shown in Tab. 3, these condition streams significantly impact our model's performance. Without the visible mask feature concatenated, performance drops substantially, indicating that the model struggles to identify the object of interest even with the guidance of the text prompt. This difficulty arises because the visible mask provides essential localization cues for the target object. Furthermore, incorporating the text prompt condition stream also results in a notable performance improvement. This enhancement is likely because the text description aids the model in reason-

Table 3. **Effect of Condition Streams.** The two conditional streams (i.e. visible mask, text prompt) play significantly roles in improving generation quality and temporal consistency .

Visible Mask	Text Prompt	Easy Cases			Hard Cases		
		CLIP↑	LSIPS↓	TC↑	CLIP↑	LSIPS↓	TC↑
✗	✓	0.76	0.52	0.81	0.74	0.56	0.79
✓	✗	0.90	0.09	0.94	0.90	0.18	0.89
✓	✓	0.93	0.06	0.96	0.92	0.14	0.93

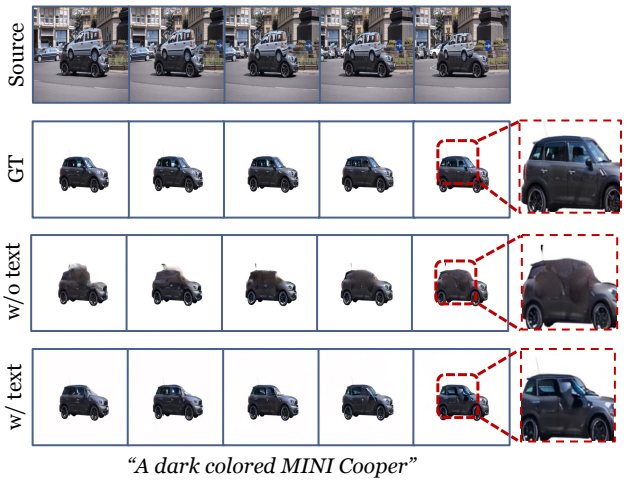

Figure 6. Qualitative comparison with and without the guided text description.

Table 4. **Effect of temporal diffusion stride for long video inference.** Smaller strides ($o = 2$ or $o = 4$) achieve the best high-level similarity, low-level detail, and temporal consistency, while larger strides or no overlap lead to reduced performance.

TD Stride	CLIP↑	LSIPS↓	TC↑
$o = 2$	0.91	0.10	0.94
$o = 4$	0.92	0.10	0.94
$o = 8$	0.88	0.11	0.87
$o = N'$	0.87	0.11	0.86

ing about the object, particularly in scenarios where the object is heavily occluded. In challenging cases, the model's performance with the text prompt is markedly better than without it, demonstrating the text prompt's critical role in guiding object recognition. Figure 6 demonstrates the advantages of the text-conditioned stream in accurately completing objects with substantial occlusion.

Effect of Temporal Diffusion for inference long videos Tab. 4 explores the effect of different temporal diffusion strides (TD Stride) on long video inference. The TD stride represents the overlap between video frames during inference, with $o = N$ indicating no overlap (i.e., no use of temporal diffusion). For strides $o = 2$ and $o = 4$, the performance is the highest, with both achieving similar CLIP scores (0.91 and 0.92, respectively), low LSIPS (0.10), and strong Temporal Consistency (TC) at 0.94. This suggests that these strides provide a good balance between maintaining high-level image similarity, low-level detail, and temporal consistency. When the stride increases to $o = 8$, the performance drops, with a CLIP score of 0.88 and a slight rise in LSIPS to 0.11, indicating a decline in high-level similarity and low-level detail. TC also drops to 0.87, showing reduced temporal stability. With no overlap ($o = N'$), the performance is the lowest, with a CLIP score of 0.87, LSIPS at 0.11, and TC at 0.86, indicating reduced overall quality and consistency. These results demonstrate that smaller strides ($o = 2$ or $o = 4$) yield the best temporal consistency and image similarity, while larger strides or no overlap result in a decline in video quality and temporal stability.

5. Conclusion & Discussion

Conclusion. In this work, we have explored the challenging problem of video amodal completion. To this end, we have introduced a large-scale synthetic dataset, equipped with detailed descriptions of objects of interest, enabling effective zero-shot transfer to natural video scenarios. Additionally, we proposed a diffusion-based, text-guided video amodal completion framework enhanced with a motion continuity module to ensure temporal consistency across video frames. By incorporating temporal diffusion, we managed long video sequences with improved inference accuracy and coherence. As the first research in video amodal completion, this work plays the role of a baseline for advancing video amodal completion and its applications in video editing, analysis, and beyond.

Discussion. Video amodal completion is an emerging field that extends the concept of image amodal completion into the video domain. While the current results mark an important milestone, the field remains in its infancy, with substantial opportunities for further exploration and refinement. Future research should focus on expanding the scope of video amodal completion to tackle more complex and challenging real-world application such as video object tracking in occlusion environments, 3D video reconstruction, virtual and augmented reality.

References

[1] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *ICRA*, pages 5085–5092. IEEE, 2022. 1

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2, 5

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 4

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

[5] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024. 2

[6] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 2

[8] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, 2018. 2

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2, 6

[10] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 4

[11] Ke Fan, Jingshi Lei, Xuelin Qian, Miaopeng Yu, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Rethinking amodal video segmentation from learning supervised signals with object-centric representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1272–1281, 2023. 2

[12] Patrick Follmann, Tobias Bottger, Philipp Hartinger, Rebecca Konig, and Markus Ulrich. Mvtec d2s: densely segmented supermarket dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 569–585, 2018. 1

[13] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *WACV*, pages 1328–1336. IEEE, 2019. 1

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[15] Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Coarse-to-fine amodal segmentation with shape prior. In *ICCV*, pages 1262–1271, 2023. 1

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[17] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 2

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1, 2, 3

[20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 4

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[22] Cheng-Yen Hsieh, Tarasha Khurana, Achal Dave, and Deva Ramanan. Tracking any object amodally. 2023. 2

[23] Yuqi Jiang, Qiankun Liu, Dongdong Chen, Lu Yuan, and Ying Fu. Animediff: Customized image generation of anime characters using diffusion model. *IEEE Transactions on Multimedia*, 2024. 2, 4, 5

[24] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *ICCV*, 2015. 2

[25] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, pages 4019–4028, 2021. 2

[26] Philip J Kellman and Thomas F Shipley. A theory of visual interpolation in object perception. *Cognitive psychology*, 23 (2):141–221, 1991. 1

[27] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4

[28] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3, 4

[30] Ke Li and Jitendra Malik. Amodal instance segmentation. In *ECCV*, pages 677–693. Springer, 2016. 1

[31] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. *NeurIPS*, 2020. 1, 2

[32] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 2

[33] Ege Ozguroglu, Ruoshi Liu, Dídac Surś, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4, 5, 6, 7, 12, 21

[34] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 4

[35] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with KINS dataset. In *CVPR*, pages 3014–3023, 2019. 1, 2

[36] Yuzhe Qin, Rui Chen, Hao Zhu, Meng Song, Jing Xu, and Hao Su. S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In *Conference on robot learning*, pages 53–65. PMLR, 2020. 1

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[39] N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *CVPR*, 2022. 2

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4, 5

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 3

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, pages 25278–25294, 2022. 2

[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

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[46] Minh Tran, Khoa Vo, Kashu Yamazaki, Arthur Fernandes, Michael Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. *arXiv preprint arXiv:2210.06323*, 2022. 1, 2

[47] Rob van Lier. Investigating global effects in visual occlusion: From a partly occluded square to the back of a tree-trunk. *Acta Psychologica*, 102(2-3):203–220, 1999. 1

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3, 5

[49] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[50] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[51] Rundi Wu, Ruoshi Liu, Carl Vondrick, and Changxi Zheng. Sin3dm: Learning a diffusion model from a single 3d textured shape. *arXiv preprint arXiv:2305.15399*, 2023. 2

[52] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9109, 2024. 1, 2, 3, 4, 6, 12, 21

[53] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 4

[54] Jian Yao, Yuxin Hong, Chiyu Wang, Tianjun Xiao, Tong He, Francesco Locatello, David P Wipf, Yanwei Fu, and Zheng Zhang. Self-supervised amodal video object segmentation. *NeurIPS*, 35:6278–6291, 2022. 1

[55] Xuyan Yun, Simon J Hazenberg, and Rob van Lier. Temporal properties of amodal completion: Influences of knowledge. *Vision Research*, 145:21–30, 2018. 1

[56] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28003–28013, 2024. 1, 2

[57] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *CVPR*, pages 3784–3792, 2020. 2, 3

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[59] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2024. 2, 5, 6

[60] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *CVPR*, 2017. 2

Text-Guided Video Amodal Completion

Supplementary Material

In this supplementary, we will delve deeper into the following aspects:

- Appendix A: We provide auxiliary qualitative results and comparison, showcasing the advantages of the proposed method.
- Appendix B We provide more experiment setup details regarding user preference study.

A. Additional Qualitative Results

A.1. Additional Zero-shot Inference on Natural Videos

Figure I to III present additional qualitative results demonstrating zero-shot performance on natural videos with real occlusion. These results highlight the robustness and adaptability of our method, producing realistic and temporally coherent amodal completions in natural, unseen video contexts. In specific, as show in Figure I, "A brownish-gray monkey running" the method successfully reconstructs the occluded parts of the monkey, producing a temporally consistent and realistic view of the animal's full shape throughout its motion. Similarly, in Figure II, "A red parrot cichlid swimming," the method exhibits perceptual realism by completing the fish's occluded body parts with smooth motion across frames, maintaining visual consistency.

A.2. Additional Qualitative Comparison

In Sec. 4.2, we compare our work with prior related works, specifically Pix2gestalt [33] and ProgressiveAmodal [52]. This section provides additional qualitative comparisons between our work and these related works, as shown in Figure IV to IX. Overall, our method showcases its advantage in delivering realistic, perceptually complete reconstructions with superior temporal alignment, which are critical for video amodal completion tasks. Notably, the reconstructed details exhibit higher realism and fidelity, maintaining the object's structure. Our approach also demonstrates remarkable robustness in challenging scenarios where objects are heavily occluded or undergo significant temporal transformations. Compared to the baselines, our method produces outputs with enhanced temporal consistency and smoother motion continuity.

For example, Figure VI showcase a goat standing in a natural setting, the ground truth once again provides a reference for ideal performance. The proposed method surpasses alternatives by preserving the goat's shape and completing occluded regions with higher perceptual realism, maintaining a consistent temporal flow across frames. In contrast, PA and Pix2gestalt introduce distortions, such as inconsistent shapes or unnatural blending of occluded areas. Similarly, in Figure VIII while competing methods struggle with temporal flickering and incomplete reconstructions of occluded details, the proposed method excels in maintaining realistic outlines, smooth transitions, and fine details like the person's arm.

B. Additional Experiment Details

B.1. User study

To measure perceptual video quality more effectively, we conducted a user preference study using a custom-designed interface. The study involved 20 university students (8 graduate students, 12 undergraduate students) from different ethnicities and academic background. Participants were tasked with evaluating the results of different amodal completion methods, focusing on how complete and realistic the generated outputs appeared.

As shown in Figure X, the interface displayed five videos side by side: the input video, ground truth, and outputs from three different methods: Pix2gestalt [33], ProgressiveAmodal [52], and our proposed approach. For each sample, the results from these methods were randomly shuffled and displayed as Method 1, Method 2, and Method 3. Each participant was instructed to review the videos and make a selection based on which generated result they found most complete and realistic. The ground truth video was provided as a reference for participants to understand the ideal completion outcome. Participants were shown a set of 20 video, evenly divided between 10 easy cases (0-50% occlusion) and 10 hard cases (50% occlusion). For each set, participants viewed the input video to understand the scene and degree of occlusion. They then observed the results from each method alongside the ground truth. Using the interface, they selected the video they considered most satisfactory in terms of perceptual completeness and realism. Participants were guided to focus on: (i) perceptual completeness: How well the occluded parts were reconstructed, ensuring consistency in appearance; (ii) realism: The natural flow and appearance of motion throughout the video; (iii) temporal consistency: The smooth transition across video frames without artifacts or noticeable disruptions. The study was conducted in a controlled environment with consistent lighting and display settings. After each selection, participants confirmed their choices by clicking the Select button below the chosen method. Results were logged automatically, ensuring randomization of video order to reduce selection bias.

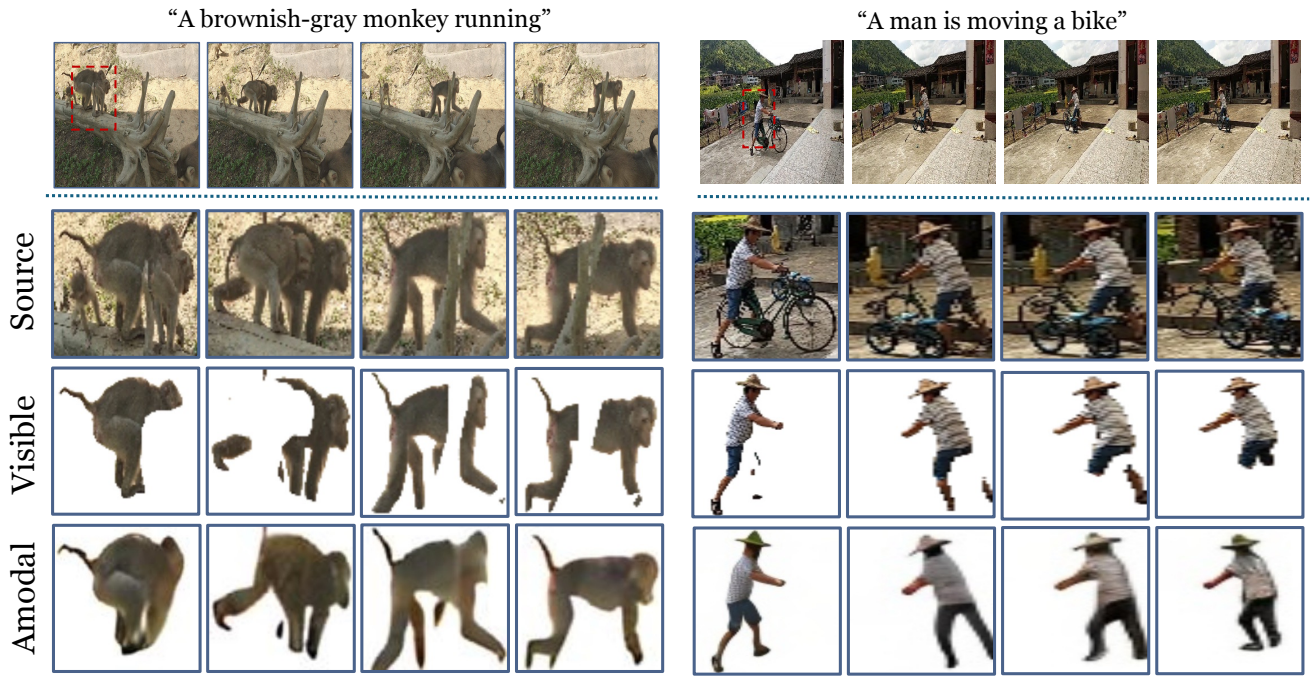

Figure I. **Additional Qualitative Results 1** showing zero-shot amodal completion on natural videos. Best viewed in zoom and color.

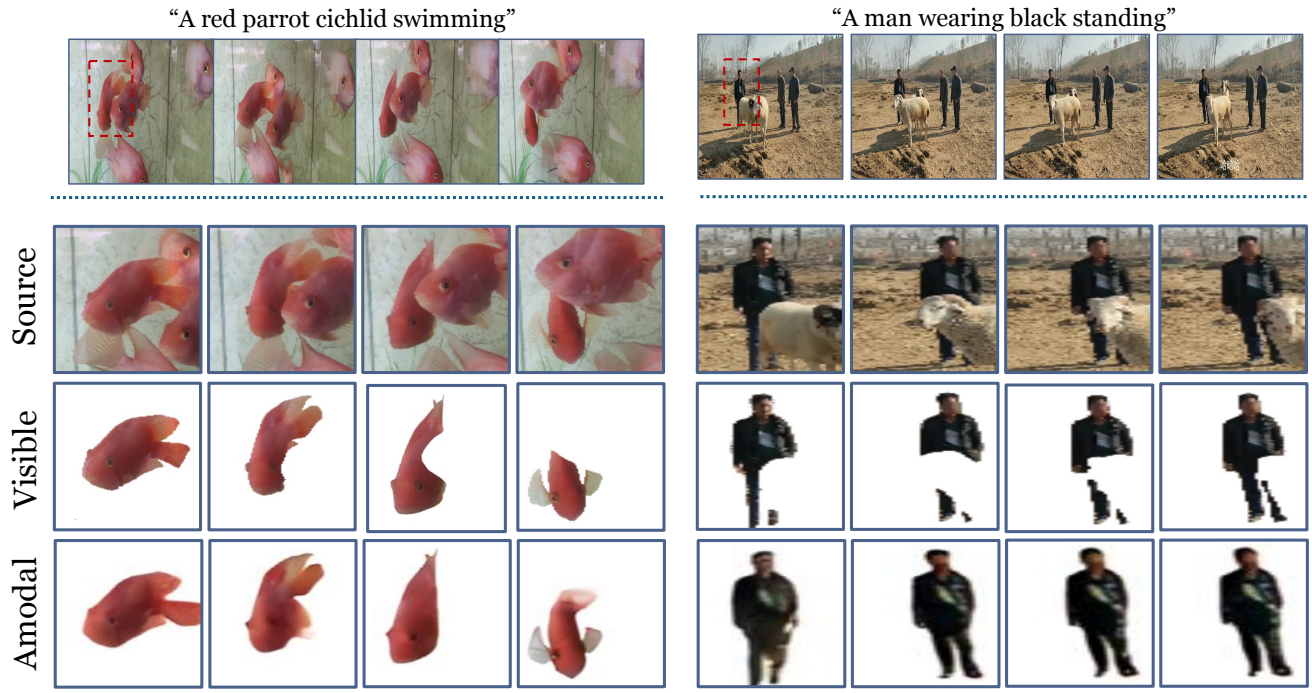

Figure II. **Additional Qualitative Results 2** showing zero-shot amodal completion on natural videos. Best viewed in zoom and color.

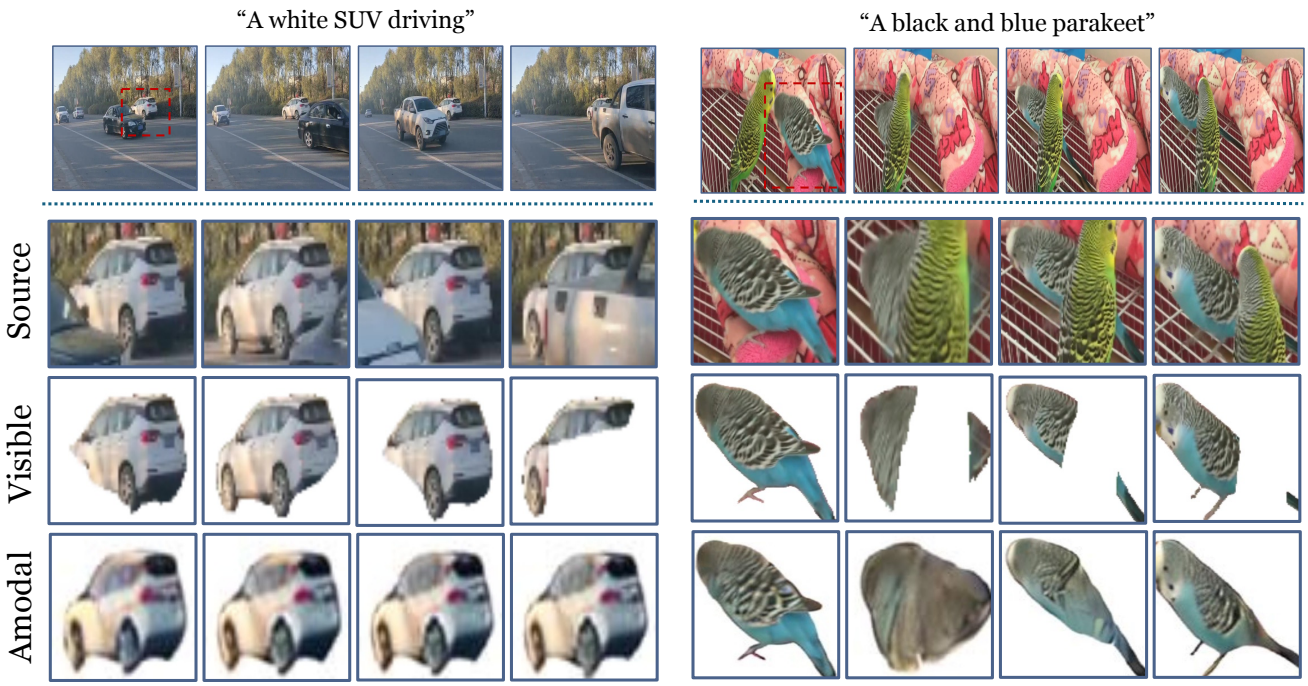

Figure III. **Additional Qualitative Results 3** showing zero-shot amodal completion on natural videos. Best viewed in zoom and color.

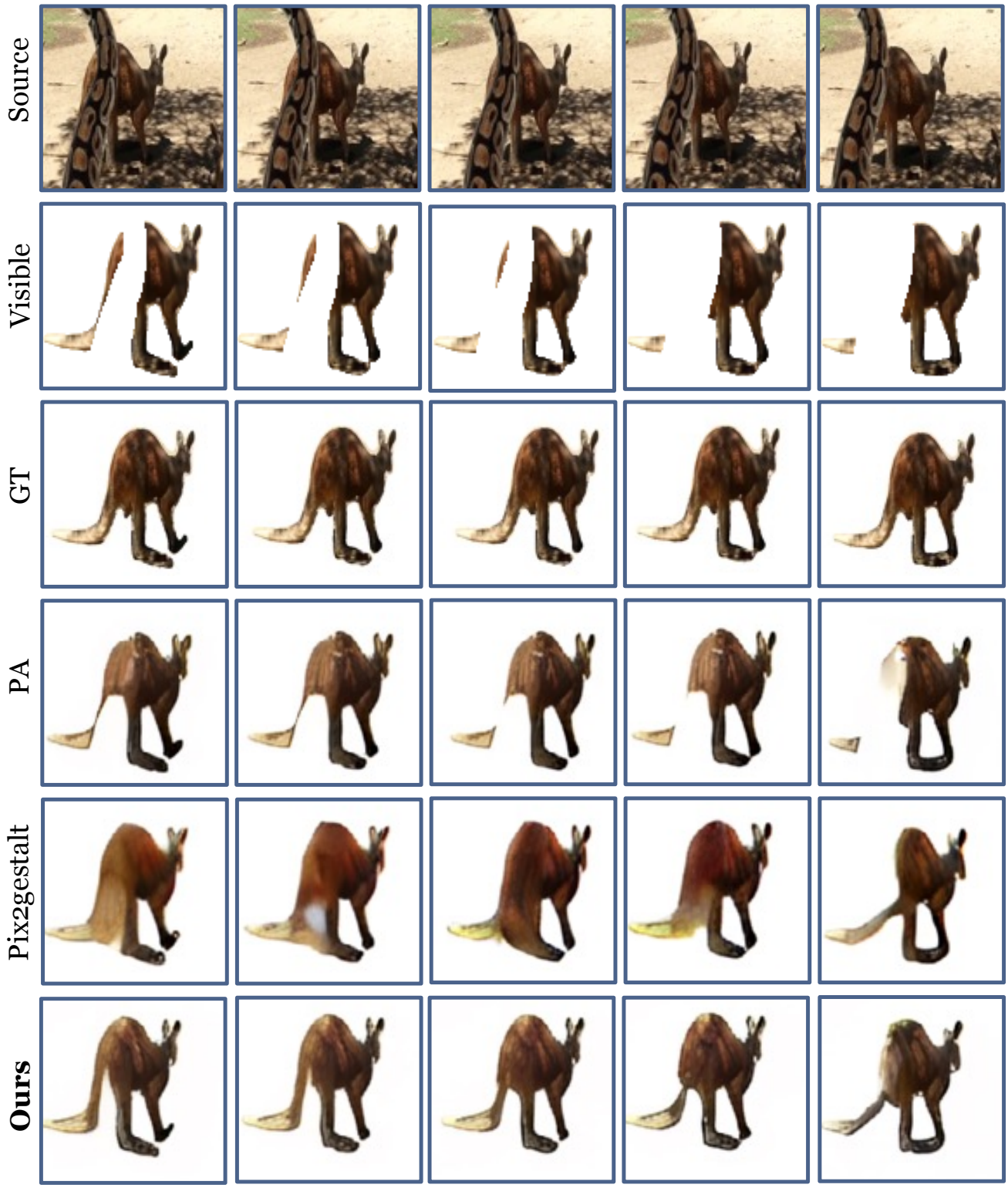

Figure IV. **Additional Qualitative Comparison 1.**

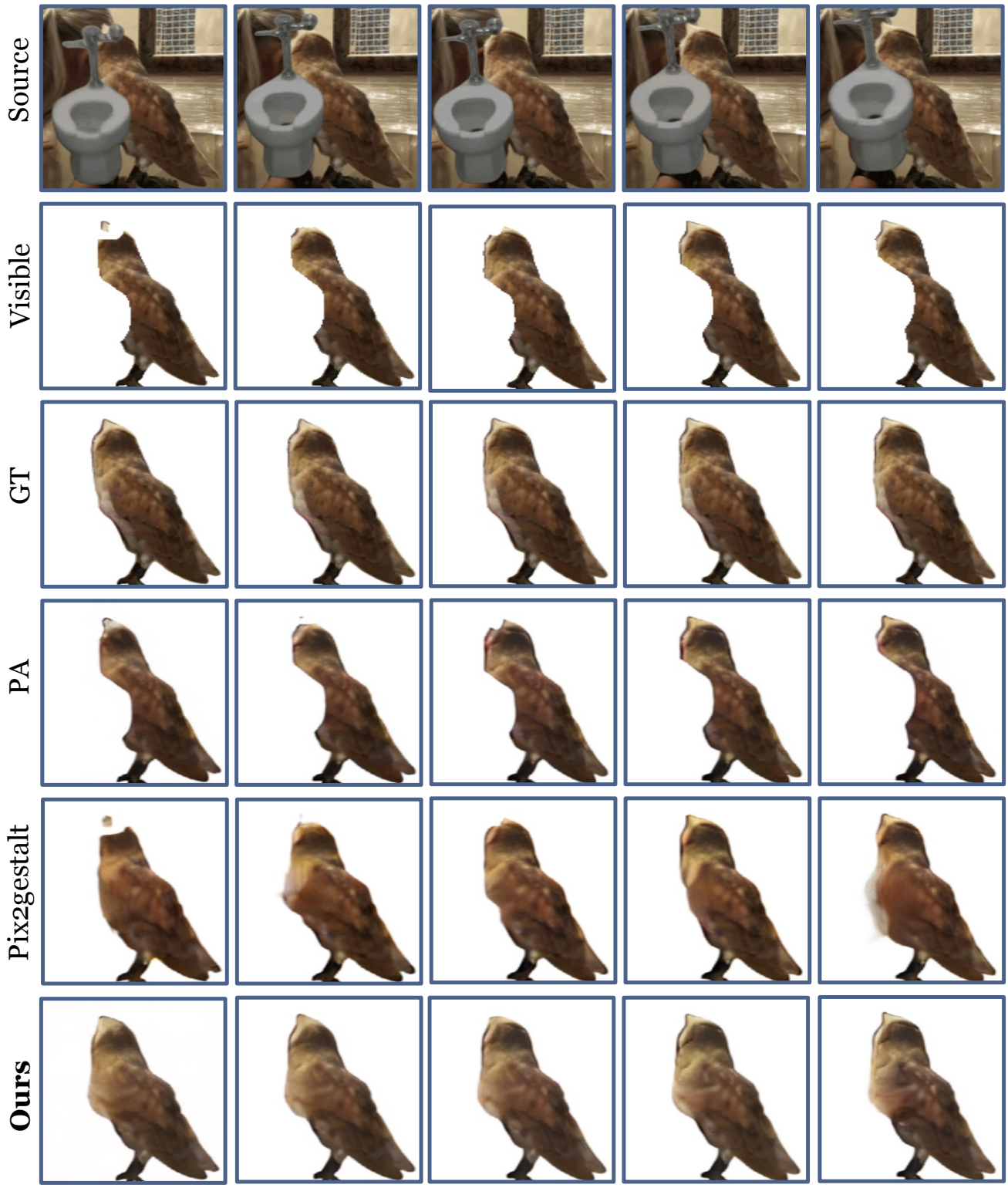

Figure V. **Additional Qualitative Comparison 2.**

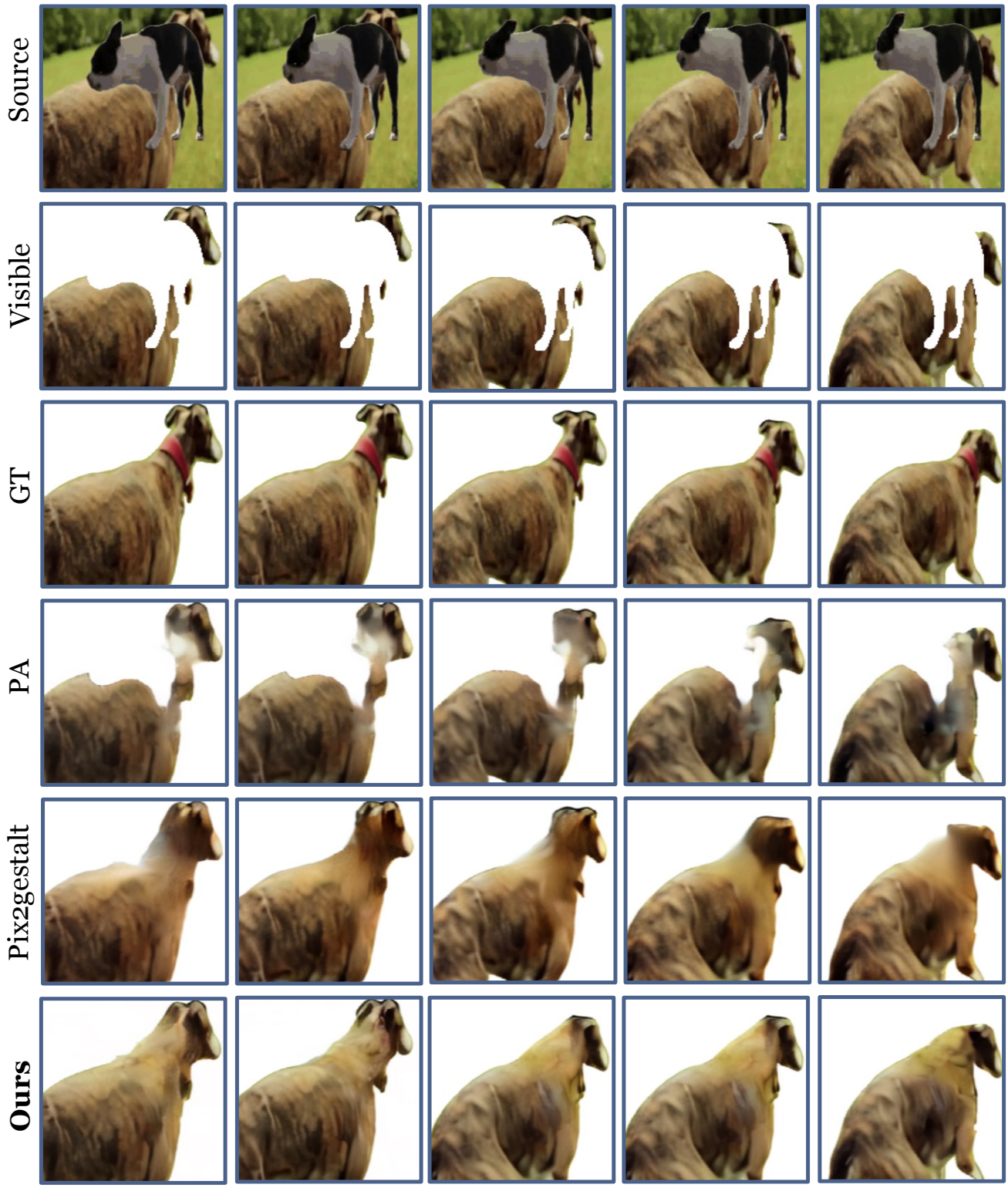

Figure VI. **Additional Qualitative Comparison 3.**

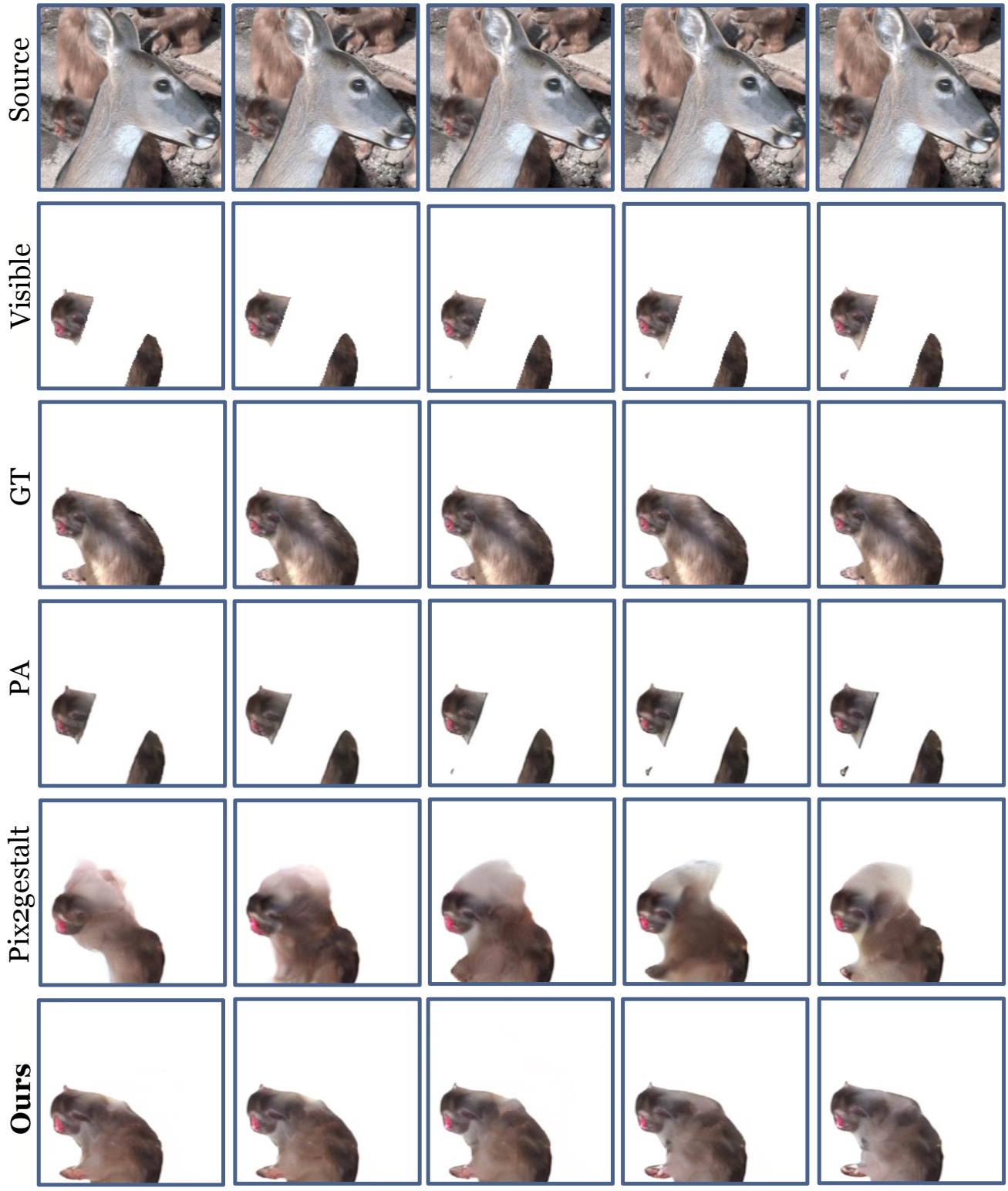

Figure VII. **Additional Qualitative Comparison 4.**

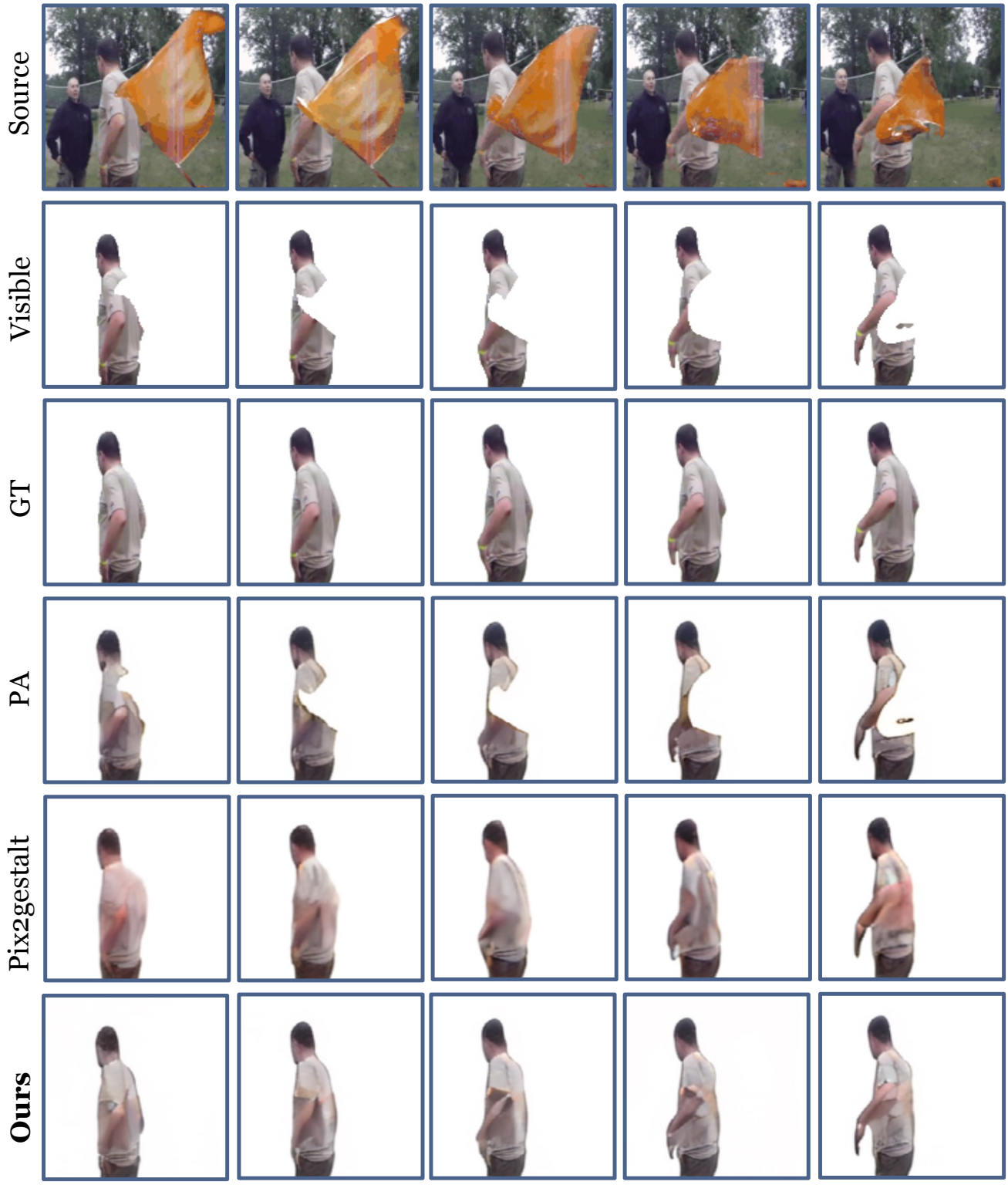

Figure VIII. **Additional Qualitative Comparison 5.**

Figure IX. **Additional Qualitative Comparison 6.**

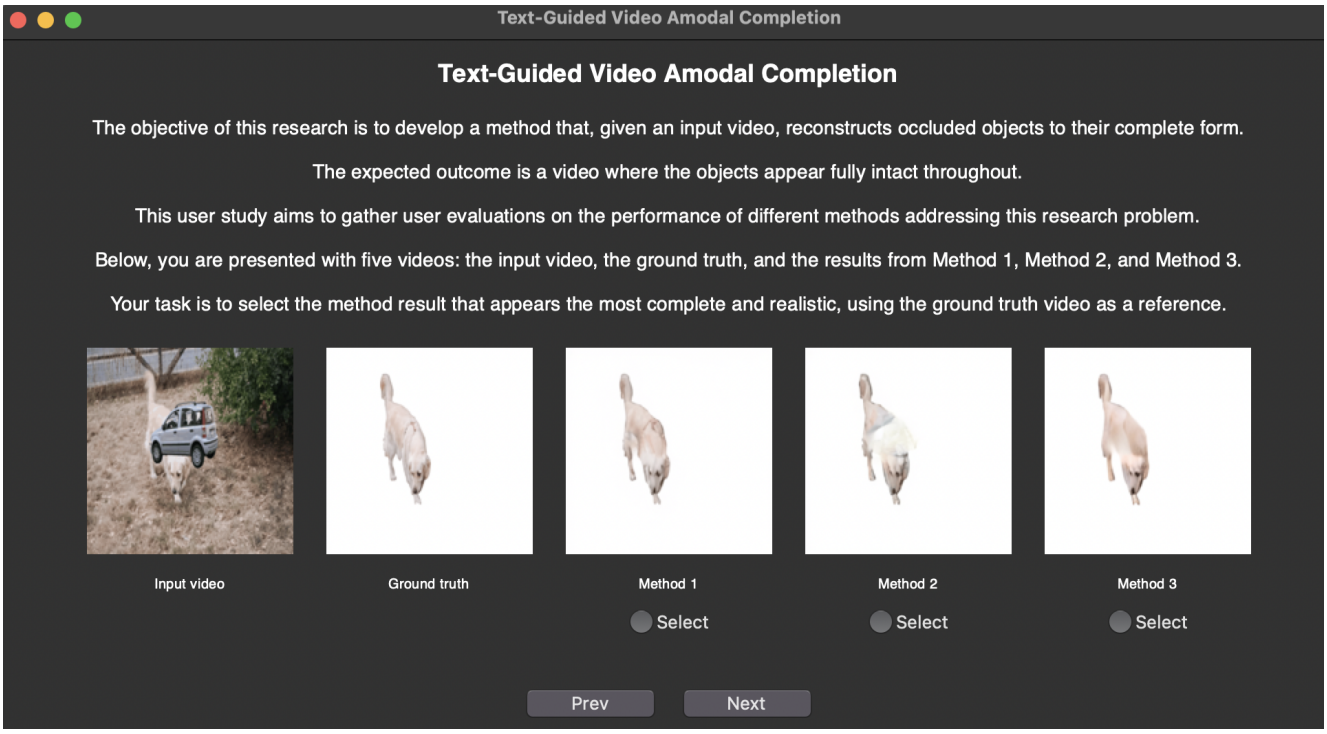

Figure X. User interface used for the user preference study. Participants were asked to select the method result that appeared most complete and realistic. The three methods evaluated were Pix2gestalt [33], ProgressiveAmodal [52], and our proposed approach. For each sample, the results from these methods were randomly shuffled and displayed as Method 1, Method 2, and Method 3.