REVEAL: Reconstructing Video Amodal Content via Language and Flow Guidance

Anonymous authors
Paper under double-blind review

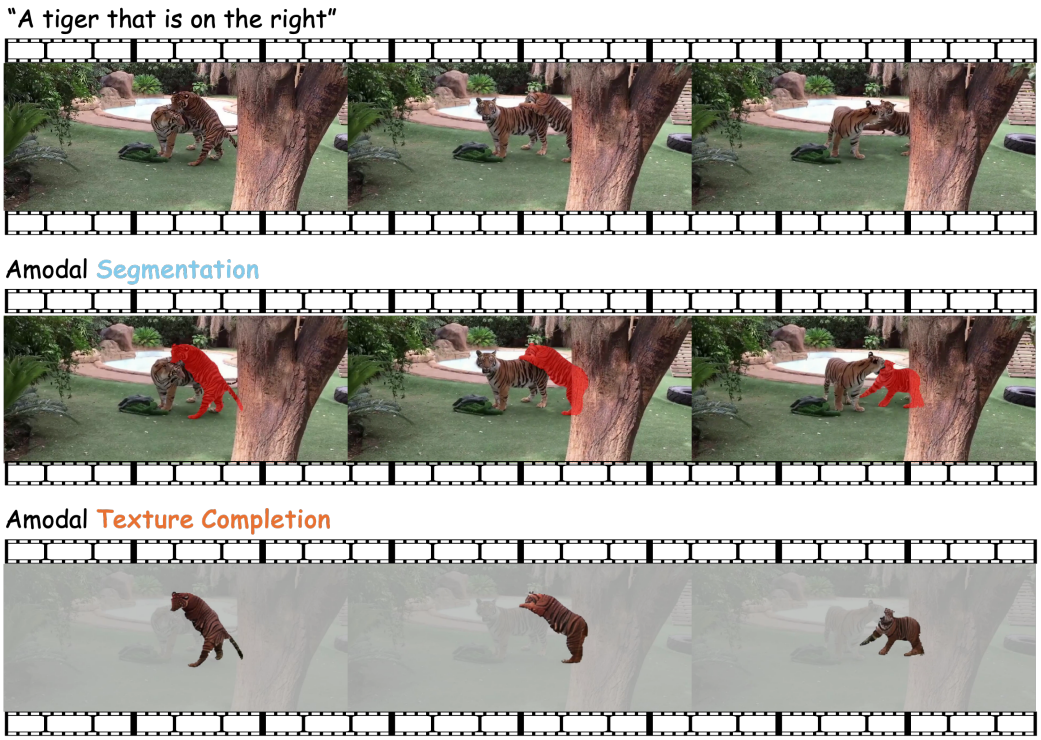

Figure 1: We tackle the problem of open-vocabulary video amodal segmentation and content completion: given an input video and a text prompt, REVEAL generates amodal (visible + invisible) masks and RGB content via a two-stage pipeline guided by optical flow and language priors.

Abstract

Amodal perception enables humans to perceive entire objects even when parts are occluded, a remarkable cognitive skill that artificial intelligence struggles to replicate. Recent diffusion-based methods have extended amodal completion from images to videos, yet they lack auxiliary priors to guide the reconstruction of heavily occluded content and to maintain temporal consistency over long sequences. We present **REVEAL** (**RE**constructing **V**id**E**o **A**modal content via **L**anguage and flow guidance), a unified diffusion-based framework that integrates complementary motion and semantic priors for video amodal completion. A user-provided text query serves a dual role: it drives open-vocabulary video segmentation to obtain visible masks, and it provides semantic guidance for texture reconstruction. For amodal segmentation, we introduce optical flow guidance: by warping visible masks from previous frames to the current frame, the flow-warped mask propagates visible content into occluded regions and approximates the object's current shape, providing a strong shape prior even under simultaneous deformation and occlusion. For texture completion, text guidance constrains the reconstruction of occluded appearance while providing a stable semantic refer-

ence that helps maintain visual coherence throughout the reconstruction. We also contribute **LAVAT**, a new benchmark featuring long video sequences paired with text descriptions, enabling evaluation of text-guided video amodal completion under heavy occlusion. Extensive experiments demonstrate that REVEAL achieves state-of-the-art performance on existing benchmarks while maintaining high-quality temporal consistency over extended sequences. The project page is available at: https://lancof.github.io.

1 Introduction

Humans possess an extraordinary ability to perceive objects as complete entities, even when parts are obscured. In everyday life, objects frequently occlude one another, yet we effortlessly infer and reconstruct their hidden regions, a capability known as amodal perception or amodal completion (Kellman & Shipley, 1991; van Lier, 1999). The human visual system relies on cues such as shape continuity and symmetry (van Lier, 1999), as well as extensive prior knowledge of the physical world (Yun et al., 2018). Replicating this cognitive process in artificial intelligence (AI) remains highly challenging, despite significant advances in computer vision.

Enabling amodal completion in AI would benefit numerous applications, including robotics (Qin et al., 2020; Back et al., 2022), autonomous driving (Qi et al., 2019), and augmented reality (Ozguroglu et al., 2024). Inspired by human perception, early research primarily focused on amodal segmentation (Li & Malik, 2016; Follmann et al., 2018; 2019; Tran et al., 2022; Yao et al., 2022; Gao et al., 2023), aiming to recover the full object mask. More recently, denoising diffusion models (Ho et al., 2020; Rombach et al., 2022) have advanced amodal content completion for images (Ozguroglu et al., 2024; Xu et al., 2024; Zhan et al., 2024) and subsequently for videos (Chen et al., 2025; Lu et al., 2025).

Despite this progress, recent video amodal completion methods often lack auxiliary priors to guide the reconstruction of occluded regions. This limitation becomes particularly critical under heavy occlusion, where inferring hidden content is inherently ambiguous or even ill-posed (Fan et al., 2023; Gao et al., 2023; Yao et al., 2022), as well as in long video sequences, where such priors are essential for maintaining temporal consistency throughout the reconstruction (Yao et al., 2022).

For video amodal segmentation, motion information is a particularly valuable source of prior knowledge. Optical flow encodes how visible regions move between frames, so warping a previous frame's mask with the estimated flow can propagate visible content into occluded areas of the current frame, providing a direct shape prior. This is especially beneficial when an object simultaneously deforms and becomes occluded: in such cases, the object's shape in earlier frames no longer matches its current shape, making appearance-based cues unreliable. Optical flow addresses this by capturing the underlying motion field, allowing the model to approximate the object's current shape even as it changes.

For amodal texture completion, the appearance of occluded regions must be reconstructed consistently across the entire video. Semantic priors, such as language descriptions, provide high-level guidance about the object's category and attributes, constraining the reconstruction of hidden content. Moreover, since the text description remains constant across all frames, it provides a stable semantic reference that helps maintain visual coherence throughout the reconstruction.

In this work, we propose REVEAL (**RE**constructing **V**id**E**o **A**modal content via **L**anguage and flow guidance). Our framework integrates complementary motion and semantic priors within a diffusion-based architecture. Instead of requiring a visible mask as input, REVEAL allows the user to provide a text query specifying the object of interest. The text query plays a dual role in our pipeline: it drives open-vocabulary video segmentation (e.g., SAM3 (Carion et al., 2025)) to obtain visible masks across frames, and it serves as a semantic prior in the texture completion stage, guiding the reconstruction of occluded content with object-class knowledge. This end-to-end integration of text makes the framework open vocabulary (Figure 1).

To incorporate motion priors, we introduce optical flow guidance. Specifically, we warp the visible mask from a previous frame to the current frame using the estimated optical flow, producing a flow-warped mask that approximates the object's full extent in the current frame. Because the warped mask can recover regions

that are currently occluded but were previously visible, it provides a strong shape prior for amodal mask prediction. Furthermore, when occlusion and non-rigid deformation occur simultaneously, the flow-warped mask captures the object's shape transformation, offering a reliable cue where static shape priors from earlier frames would fail. To incorporate semantic priors, we leverage textual description guidance. Text features encode semantic attributes and canonical shape information of the target category, constraining the reconstruction of occluded content. By combining motion consistency with semantic understanding, our framework provides stronger guidance than previous methods, helps improve performance and consistency under severe occlusions and long videos.

In addition, we present LAVAT, a new benchmark for text-guided video amodal completion. LAVAT contains long video sequences paired with text descriptions, enabling evaluation of both temporal consistency and text-video alignment. It addresses the limitations of prior benchmarks that lack text annotations and focus exclusively on short clips.

In summary, our contributions are: (1) We introduce REVEAL, a unified framework that integrates optical flow as a motion prior and language as a semantic prior for video amodal completion, addressing the lack of auxiliary guidance under heavy occlusion. (2) We introduce LAVAT, a new benchmark for text-guided video amodal completion, featuring long video sequences with text descriptions for comprehensive evaluation. Extensive experiments demonstrate that REVEAL achieves state-of-the-art performance on video amodal completion benchmarks.

2 Related Work

2.1 Amodal Completion and Segmentation

Amodal visual understanding addresses scenarios where objects are partially hidden by foreground elements. Amodal segmentation recovers the full object mask including occluded regions (Zhu et al., 2017; Ke et al., 2021; Qi et al., 2019; Reddy et al., 2022; Ling et al., 2020), while amodal detection predicts the complete object extent (Kar et al., 2015; Hsieh et al., 2023). Progress in mask-based modeling and the availability of closed-world (Zhan et al., 2020; Ling et al., 2020; Ke et al., 2021; Qi et al., 2019; Kar et al., 2015; Zhu et al., 2017) and synthetic datasets (Ehsani et al., 2018) have fueled strong methods such as PCNet (Zhan et al., 2020) and AISFormer (Tran et al., 2022). More recently, amodal completion has emerged, aiming to reconstruct the entire object shape (Ehsani et al., 2018; Zhan et al., 2020; Ozguroglu et al., 2024). Pix2gestalt (Ozguroglu et al., 2024) fine-tunes a large diffusion model on synthetic data, while MCAmodal (Xu et al., 2024) and OpenAmodal (Ao et al., 2025) reframe the task as inpainting and leverage pretrained models directly. Building on image-level advances (Ozguroglu et al., 2024; Xu et al., 2024; Zhan et al., 2024), recent efforts extend amodal completion to videos (Chen et al., 2025; Lu et al., 2025). *However, existing video methods lack sufficient auxiliary cues for recovering severely occluded regions and maintaining temporal consistency over longer sequences. Our REVEAL addresses these limitations by incorporating motion and semantic priors.*

2.2 Video Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) have emerged as powerful generative methods, applied to image generation (Ho & Salimans, 2022; Song et al., 2021; Dhariwal & Nichol, 2021), video generation (Blattmann et al., 2023a; Ho et al., 2022a; Gu et al., 2023), motion generation (Guo et al., 2024; Zhang et al., 2024), and 3D generation (Liu et al., 2023; Deitke et al., 2023; Wu et al., 2024). A key milestone was (Dhariwal & Nichol, 2021), showing diffusion models can surpass GANs (Goodfellow et al., 2014) in image synthesis, followed by Stable Diffusion (Rombach et al., 2022) which enables text-guided generation and editing (Brooks et al., 2023; Gal et al., 2022; Ruiz et al., 2023). For video synthesis, VDM (Ho et al., 2022b) extends text-to-image diffusion to joint image-video training, Imagen Video (Ho et al., 2022a) employs cascaded diffusion with super-resolution, and Make-A-Video (Singer et al., 2022) leverages unsupervised video data. Gen-1 (Esser et al., 2023) introduces structure-guided video editing, while Tune-A-Video (Wu et al., 2023) enables one-shot generation. Recent models such as CogVideoX (Yang et al., 2024b) and WAN-Video (Wan et al., 2025) support diverse tasks including text-to-video, image-to-video, and video editing. A complementary line of work conditions these backbones on motion or addresses long-horizon

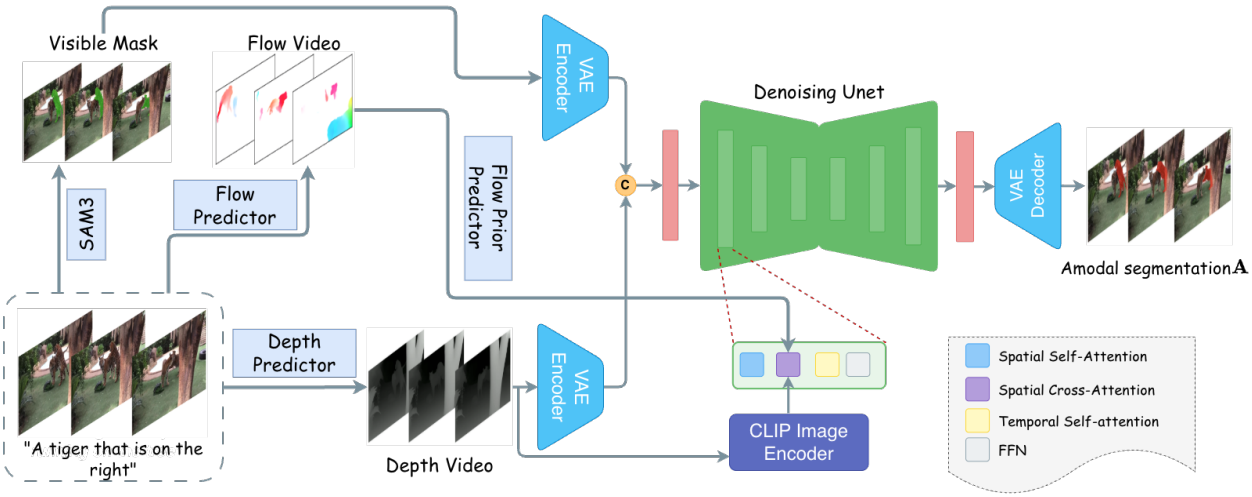

Figure 2: **Video Amodal Mask Segmentation Stage.** The model uses optical flow as a motion prior, warping visible masks from previous frames to propagate shape information into occluded regions.

generation: MOFA-Video (Niu et al., 2024) warps multi-scale first-frame features under a dense flow field through a ControlNet-style adapter for controllable image-to-video animation, and Self-Forcing (Huang et al., 2025) mitigates exposure bias in autoregressive video diffusion via post-training distillation on a causal-DiT backbone. *Our approach builds upon video diffusion models to leverage shape and texture priors learned from large-scale pretraining.*

3 Methodology

Problem Setting. Let $\mathbf{I} \in \mathbb{R}^{N\times 3\times H\times W}$ denote a video sequence and let $\mathbf{p}$ be an initial text prompt (e.g., "a golden retriever") specifying the object of interest in the first frame. We first employ an open-vocabulary video segmentation model (e.g., SAM3 (Carion et al., 2025)) to obtain the visible masks $\mathbf{V} \in \mathbb{R}^{N\times H\times W}$ of the object across the entire sequence. Our framework consists of two stages. In the first stage, we predict amodal masks $\mathbf{A} \in \mathbb{R}^{N\times H\times W}$ that capture the complete spatial extent of the object, including occluded regions. In the second stage, conditioned on $\mathbf{A}$, we reconstruct the RGB content (texture) of the occluded regions, yielding $\mathbf{C} \in \mathbb{R}^{N\times 3\times H\times W}$. The objective is for the visible portions of $\mathbf{C}$ to be faithful to the corresponding regions in the input video $\mathbf{I}$, while the completed (occluded) regions should blend naturally with the scene, preserving contextual coherence and avoiding physically implausible object structures.

We adopt a two-stage design because it provides greater stability and consistency. A single-stage approach often struggles to infer complete object shapes, making it even more challenging to generate coherent texture for the occluded regions (Chen et al., 2025).

3.1 REVEAL

Our proposed REVEAL includes two sequential stages: *Video Amodal Mask Segmentation with Optical Flow Guidance*, and *Video Amodal Texture Completion with Text Guidance*. The two stages are equipped with prior knowledge to address occlusion. Specifically, in the first stage, optical flow serves as a motion prior, warping visible masks from previous frames to propagate shape information into occluded regions. In the second stage, text guidance provides a semantic prior, leveraging pretrained diffusion models to supply object-class information. These textual features encode canonical shapes and properties, enabling more accurate reconstruction of occluded content.

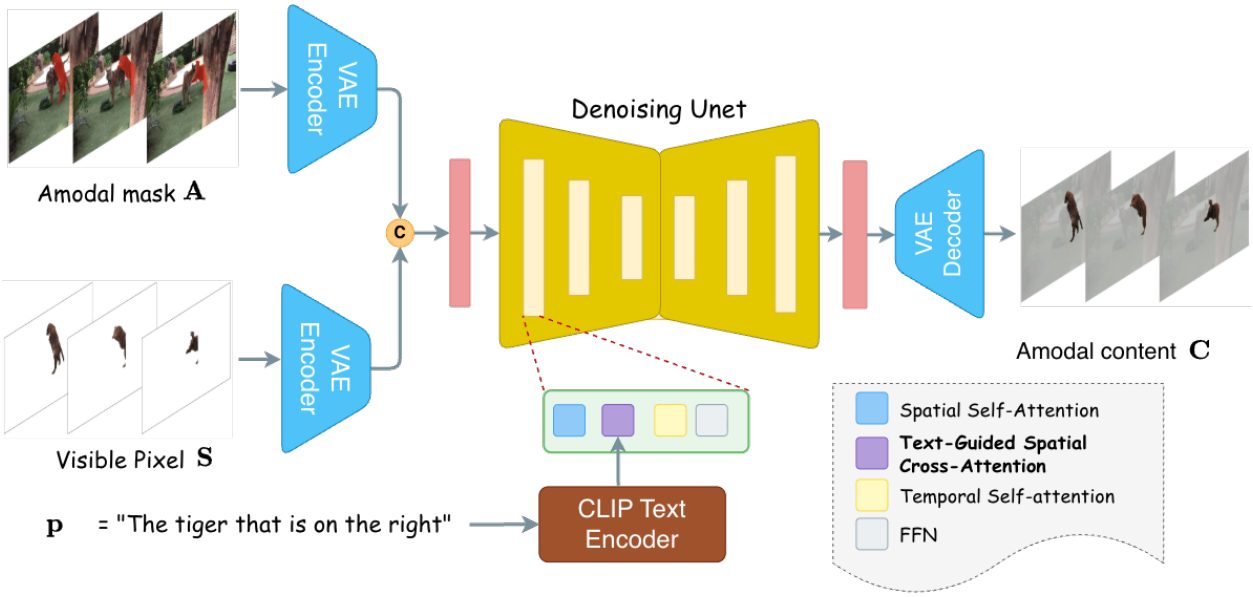

Figure 3: **Video Amodal Texture Completion Stage.** The second stage conducts amodal texture completion with text guidance as semantic priors, leveraging pretrained diffusion models to reconstruct accurate content based on canonical object shapes and properties.

3.1.1 Stage 1: Video Amodal Mask Segmentation with Optical Flow Guidance

We first train the *Video Amodal Mask Segmentation with Optical Flow Guidance* module to predict the video amodal segmentation masks $\mathbf{A}$. Given an RGB video $\mathbf{I}$ and its corresponding visible mask $\mathbf{V}$ for the object of interest, this module aims to predict the amodal segmentation masks $\mathbf{A}$ that capture both visible and occluded regions of the object. Our approach builds upon a video diffusion model (e.g., SVD (Blattmann et al., 2023a)) to leverage shape and content priors learned from large-scale pretraining. Specifically, the visible masks $\mathbf{V}$ of shape $\mathbb{R}^{N\times H\times W}$ are first passed through a variational autoencoder (VAE) (Kingma & Welling, 2013). Since the VAE requires three input channels, we replicate the binary mask across channels, following the single-channel VAE adaptation in (Ke et al., 2024; Chen et al., 2025). Encoding through the VAE produces a latent tensor of shape $\mathbb{R}^{N\times C_1\times \frac{H}{s}\times \frac{W}{s}}$, where C_1 is the latent channel dimension and s is the VAE spatial downsampling factor. This latent, concatenated with a noise tensor of the same shape, is then processed by our spatio-temporal 3D U-Net (Blattmann et al., 2023b), resulting in a final input of shape $\mathbb{R}^{N\times 2C_1\times \frac{H}{s}\times \frac{W}{s}}$. Additionally, we incorporate CLIP embeddings (Radford et al., 2021) of the visible masks $\mathbf{V}$ and inject them into the transformer layers via cross-attention. This provides temporal information about the visibility of objects across surrounding frames. Inspired by (Chen et al., 2025), we further introduce pseudo-depth maps as an additional input in place of raw RGB frames. Occlusions are often caused by objects closer to the camera, and pseudo-depth offers stronger implicit cues about potential occluders than RGB images, making it a more effective indicator for identifying regions to complete (Chen et al., 2025). To obtain these pseudo-depth maps, we also employ the Depth Anything V2 monocular depth estimator (Yang et al., 2024a), which converts RGB frames into pseudo-depth maps $\mathbf{D} \in \mathbb{R}^{N\times 3\times H\times W}$ that are concatenated with the visible mask latents as extra channels in the diffusion model.

Finally, we integrate motion cues from visible regions by incorporating optical flow vectors. By warping a previous frame's visible mask to the current frame, optical flow can propagate visible content into currently occluded areas, providing a direct shape prior for amodal mask prediction. To extract optical flow, we adopt the state-of-the-art DPFlow model (Morimitsu et al., 2025), which estimates $\mathbf{F} = \{\mathbf{F}_i\}_{i=0}^{N-1}$, where $\mathbf{F}_i$ is the flow between frames $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$.

Thus, for each frame, we produce a flow prior feature as guidance for the amodal segmentation mask module. For a specific frame with visible mask $\mathbf{V}_k$, we compute $\tilde{\mathbf{V}}_{k-1\to k}$, which is the warped mask from the previous

frame to the current time step. Because the warped mask can recover regions that are currently occluded but were previously visible, it provides a strong motion prior for amodal mask prediction. Moreover, when occlusion and non-rigid deformation occur simultaneously, the flow-warped mask captures the object's shape transformation, offering a reliable cue where static shape priors from earlier frames would fail. In other words, this flow prior targets *dynamic* occlusion, an object occluded while moving or deforming. We then design a simple Flow Prior encoder g that encodes the concatenation of the two masks to $\tilde{\mathbf{F}}_{\text{prior}} \in \mathbb{R}^{N\times d}$, where d is the feature dimension. In practice, g is a lightweight CNN ($\approx$ 640K parameters): three Conv - SiLU stride-2 blocks ($6 \to 64 \to 128 \to 256$ channels) followed by global average pooling and a linear projection to d, producing one token per frame. The process can be formulated as follows:

$$\begin{aligned} \tilde{\mathbf{V}}_{k-1\to k} &= \text{warp}(\mathbf{V}_{k-1}, \mathbf{F}_{k-1\to k}), \\ \tilde{\mathbf{F}}_{\text{prior}} &= g\big([\tilde{\mathbf{V}}_{k-1\to k}, \mathbf{V}_k]\big) \end{aligned} \tag{1}$$

We then inject $\tilde{\mathbf{F}}_{\text{prior}}$ into the transformer through cross-attention, allowing the model to leverage the flow prior context. Formally, the objective of the denoising model in this stage is:

$$\min_{\theta_m} \mathbb{E}\big\|\epsilon - \epsilon_{\theta_m}(\mathbf{z}_t \oplus \mathcal{E}(\mathbf{V}) \oplus \mathcal{E}(\mathbf{D}), \mathcal{C}(\mathbf{V}), \tilde{\mathbf{F}}_{\text{prior}}, t)\big\|_2^2 \tag{2}$$

where ϵ is the sampled Gaussian noise, θ_m denotes the learnable parameters of the noise prediction network ϵ_{θ_m} for video amodal segmentation, $\mathbf{z}_t$ is the noisy latent at diffusion timestep t, $\mathcal{E}(\cdot)$ represents the VAE encoder for visible mask $\mathbf{V}$ and pseudo-depth $\mathbf{D}$, $\mathcal{C}(\cdot)$ denotes the CLIP embedding encoder for visible masks, $\tilde{\mathbf{F}}_{\text{prior}}$ denotes the flow prior feature, $\oplus$ indicates channel-wise concatenation, and $\|\cdot\|_2^2$ is the squared ℓ_2 norm. At inference, the trained θ_m is used in DDIM sampling (Song et al., 2021) to obtain the predicted video amodal masks $\mathbf{A}$.

3.1.2 Stage 2: Video Amodal Texture Completion with Text Guidance

After obtaining the video amodal masks $\mathbf{A}$ in the previous module, we input the amodal masks alongside the visible texture of the object of interest $\mathbf{S} = \mathbf{I} \odot \mathbf{V}$. Besides these two inputs, we propose a text guidance mechanism for this module. The text guidance acts as a semantic prior for completing the occluded texture. Utilizing the rich textual representations from pretrained diffusion models, text guidance supplies object-class information that informs the appearance of occluded regions. These textual features encode canonical object properties, enabling more accurate reconstruction of hidden content through semantic understanding of the object category. In detail, for the conditional latents, we encode both the amodal masks and visible pixels through the VAE to obtain latent vectors. These latents, concatenated with a noise latent, form the input of shape $\mathbb{R}^{N\times 2C_1\times \frac{H}{s}\times \frac{W}{s}}$. Additionally, we incorporate CLIP embeddings of the visible pixels $\mathbf{S}$ for cross-attention, which introduces frame-specific appearance cues.

For the text guidance, we use the CLIP encoder to process the text prompt $\mathbf{p}$ and obtain text embeddings. These embeddings are injected into the transformer via cross-attention, propagating semantic priors across frames for content completion. Because the text prompt remains constant across all frames, it provides a stable semantic reference that helps maintain visual coherence throughout the reconstruction. By combining visual cues from $\mathbf{S}$ and semantic priors from $\mathbf{p}$, the model gains both appearance-specific and class-level knowledge, leading to more precise and consistent reconstructions. Formally, the objective for this stage is:

$$\min_{\theta_c} \mathbb{E}\big\|\epsilon - \epsilon_{\theta_c}(\mathbf{z}_t \oplus \mathcal{E}(\mathbf{A}) \oplus \mathcal{E}(\mathbf{S}), \mathcal{C}(\mathbf{S}), \mathcal{C}(\mathbf{p}), t)\big\|_2^2 \tag{3}$$

where θ_c are the parameters of the denoising model for content completion, $\mathbf{z}_t$ is the noisy latent at timestep t, $\mathcal{E}(\cdot)$ denotes the VAE encoder applied to the amodal mask $\mathbf{A}$ and visible texture $\mathbf{S}$, $\oplus$ indicates channel-wise concatenation, $\mathcal{C}(\mathbf{S})$ is the CLIP embedding of the visible pixels, $\mathcal{C}(\mathbf{p})$ is the CLIP embedding of the text prompt, and $\|\cdot\|_2^2$ is the squared ℓ_2 distance between the predicted and true noise. In inference, the trained θ_c is used in DDIM sampling (Song et al., 2021) to obtain the video amodal content $\mathbf{C}$.

Table 1: Comparison of statistics with the existing video amodal completion datasets. For LAVAT, we report values for the synthetic subset (S) and real subset (R). # denotes "number of".

	Kubric	TACO-OvO	LAVAT (S)	LAVAT (R)
Video length	Short	Short	Long	Long
· Min #frames	24	14	91	41
· Max #frames	24	77	1595	661
· Mean #frames	24	19	257	151
# videos	690	2049	2,658	100
# total frames	9,660	27,342	684,037	15,139
Occlusion rate (%)	22.96	46.67	46.83	-
Text description	✗	✗	✓	✓

Long Video Inference. Although current video generative models (Blattmann et al., 2023a; Yang et al., 2024b; Wan et al., 2025; Chen et al., 2025; Lu et al., 2025) and our REVEAL can theoretically process videos of arbitrary length, significant degradation in completion quality occurs when generating videos whose total length exceeds N (Zhang et al., 2024). We thus apply Temporal Generation, the long video inference mechanism in (Zhang et al., 2024), where a sliding window of size N is introduced that incrementally synthesizes future frames based on previously generated outputs.

3.2 LAVAT Benchmark

We introduce LAVAT, a benchmark for text-guided video amodal completion that features long video sequences paired with text descriptions. LAVAT includes two subsets: (i) LAVAT-S, a large-scale synthetic-occlusion benchmark constructed by compositing objects to create controlled occlusions with amodal ground truth; and (ii) LAVAT-R, a small real-occlusion subset of 100 videos, where each video is provided with a human-annotated visible segmentation mask of the object of interest and a corresponding human-written text hint. Existing benchmarks largely overlook long-horizon evaluation and typically contain only short clips, e.g., 24 frames in Kubric (Greff et al., 2022) and 14 frames in TACO-OvO (Lu et al., 2025). Moreover, none of them provide text descriptions; since amodal completion is a generative task, text descriptions enable complementary evaluation via text–video alignment metrics in addition to pixel-level similarity.

To construct LAVAT-S, we source videos from multiple video object segmentation datasets, including DAVIS (Pont-Tuset et al., 2017), YouTube-VOS (Xu et al., 2018), and LVOS (Hong et al., 2023). These sources provide medium-to-long sequences with diverse object types and motion patterns, and often contain a single target object, which simplifies selecting unoccluded object instances to serve as amodal ground truth. We isolate object pixels using the provided masks and synthesize occluded inputs by compositing objects to create controlled occlusion scenarios, while retaining the original unoccluded object pixels as ground-truth completion targets. For text descriptions, we caption a representative frame (the first frame of the ground-truth video) using a Qwen-VL model (QWEN2.5/3) (Bai et al., 2025) with a structured instruction that enforces a compact description of the target object. The model is prompted to output a JSON object containing (i) a short noun phrase for the object category, (ii) visible attributes (e.g., color, material, shape), and (iii) the action if applicable. We then convert this structured output into a concise sentence (object + attributes + action) and perform manual refinement by human annotators for correctness. Table 1 summarizes LAVAT and compares it with Kubric (Greff et al., 2022) and TACO-OvO (Lu et al., 2025). Compared with Kubric and TACO, LAVAT exhibits wider range of video lengths with substantial long video sequences. While Kubric and TACO contain only short clips with maximum lengths of 24 and 77 frames respectively, LAVAT supports long sequences, with up to 1,595 frames in the synthetic subset and 661 frames in the real subset. The average sequence length in LAVAT, 257 frames for synthetic and 151 frames for real, is significantly higher than Kubric, 24 frames, and TACO, 19 frames, demonstrating its suitability for long-term temporal reasoning. In terms of scale, LAVAT is also considerably larger. The synthetic subset contains 2,658 videos and 684,037 frames, exceeding Kubric, 690 videos and 9,660 frames, and TACO, 2,049 videos and 27,342 frames, by a large margin. The real subset further contributes 100 videos and 15,139 frames, providing additional real-world diversity. Moreover, LAVAT exhibits a high occlusion rate of 46.83 percent, comparable to TACO at 46.67 percent and significantly higher than Kubric at 22.96 percent, making it

Table 2: Quantitative comparison with state-of-the-art methods. Bold illustrate best performance.

Dataset	Method	Texture Completion				Segmentation	Temporal Consistency	
		PSNR(↑)	SSIM(↑)	LPIPS(↓)	CLIP-T(↑)	mIoU(↑)	CLIP-F(↑)	FVD(↓)
Kubric	Pix2Video	14.876	0.704	0.325	-	27.45	0.939	1073.13
	E2FGVI	14.406	0.688	0.227	-	54.05	0.967	732.30
	pix2gestalt	20.096	0.826	0.129	-	71.95	0.911	405.89
	PCNet	20.111	0.861	0.111	-	79.70	0.958	253.77
	Diff-VAS	21.213	0.852	0.099	-	81.05	0.974	229.04
	TACO	23.509	0.889	0.077	-	80.65	0.970	186.10
	REVEAL (Ours)	**29.916**	**0.921**	**0.051**	-	**84.12**	**0.975**	**122.66**
TACO-OvO	Pix2Video	12.231	0.672	0.385	-	38.65	0.941	1638.49
	E2FGVI	13.123	0.652	0.326	-	59.65	0.966	1137.89
	pix2gestalt	19.442	0.803	0.133	-	86.30	0.920	532.01
	TACO	24.342	0.882	0.074	-	94.00	0.978	216.93
	REVEAL (Ours)	**26.881**	**0.899**	**0.068**	-	**95.12**	**0.988**	**213.77**
LAVAT-S (Ours)	pix2gestalt	21.057	0.952	0.172	0.279	83.83	0.914	555.95
	VACE	23.541	0.943	0.106	0.270	59.09	0.892	960.14
	Diff-VAS	23.933	0.948	0.081	0.271	86.60	0.935	443.59
	TACO	24.821	0.951	0.082	0.272	85.28	0.925	471.42
	REVEAL (Ours)	**27.473**	**0.963**	**0.073**	**0.285**	**90.29**	**0.945**	**310.90**

Table 3: Evaluation on LAVAT-R (real occlusions). We report text-video alignment (CLIP-T) and human preference (%) from a user study.

Method	CLIP-T(↑)	User Pref. (↑)
Pix2gestalt Ozguroglu et al. (2024)	0.274	6%
Diff-VAS Chen et al. (2025)	0.279	12%
TACO Lu et al. (2025)	0.280	11%
REVEAL (Ours)	**0.298**	**71%**

more challenging and realistic for amodal completion tasks. Importantly, unlike prior datasets, LAVAT provides textual descriptions for both synthetic and real subsets, enabling multimodal learning and text-guided amodal completion, which is not supported in Kubric or TACO.

4 Experimental Results

4.1 Experimental Setup

Implementation Details. For training, we follow a similar training setting to Diff-VAS (Chen et al., 2025) for fair comparison. We initialize from the official SVD-xt (Blattmann et al., 2023a) pretrained checkpoint and finetune on curated SAILVOS dataset from (Chen et al., 2025). We adopt the AdamW optimizer with a learning rate of 3×10^{-6}, a batch size of 8, and a frame resolution of 128×256. Training requires approximately 24 hours on 8 NVIDIA RTX A6000 GPUs. During inference, we use 25 DDIM (Song et al., 2021) denoising steps with a guidance scale of 4.5. We apply a sliding window of size 25 with a stride of 4, following (Zhang et al., 2024).

Datasets. We evaluate on three benchmarks: our LAVAT, Kubric (Greff et al., 2022), and TACO-OvO (Lu et al., 2025). To enable text guidance on Kubric and TACO-OvO, we generate text descriptions for the amodal texture ground truth using QWEN2.5 (Bai et al., 2025), followed by manual refinement.

Metrics. We evaluate texture completion using PSNR (Gonzalez, 2009), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) for image-level alignment. For LAVAT, we additionally report CLIP-T (Radford et al., 2021) to assess text–video alignment. Amodal segmentation quality is measured by mIoU. For temporal

Figure 4: Qualitative comparison between TACO (Lu et al., 2025), Diff-VAS (Chen et al., 2025), and our REVEAL on LAVAT-R benchmark.

consistency, we use CLIP-F (Esser et al., 2023) for frame-to-frame similarity and FVD (Unterthiner et al., 2018) for sequence-level coherence. For LAVAT-R, we report CLIP-T and human preference (%) from a user study.

4.2 Comparison to State-of-the-art

We compare against recent video amodal completion methods Diff-VAS (Chen et al., 2025) and TACO (Lu et al., 2025), image-level amodal completion models Pix2Gestalt (Ozguroglu et al., 2024) and PCNet (Zhan et al., 2020) extended frame-by-frame, video inpainting methods E2FGVI (Li et al., 2022) and VACE (Jiang et al., 2025), and a video editing approach Pix2Video (Ceylan et al., 2023).

Table 4: **Flow prior ablation across occlusion difficulty on LAVAT.** We report amodal mIoU (↑).

Setting	< 20%	20–40%	40–65%	> 65%
w/o flow prior	91.88	87.97	85.75	87.79
w/ flow prior	92.93	90.72	88.38	89.13

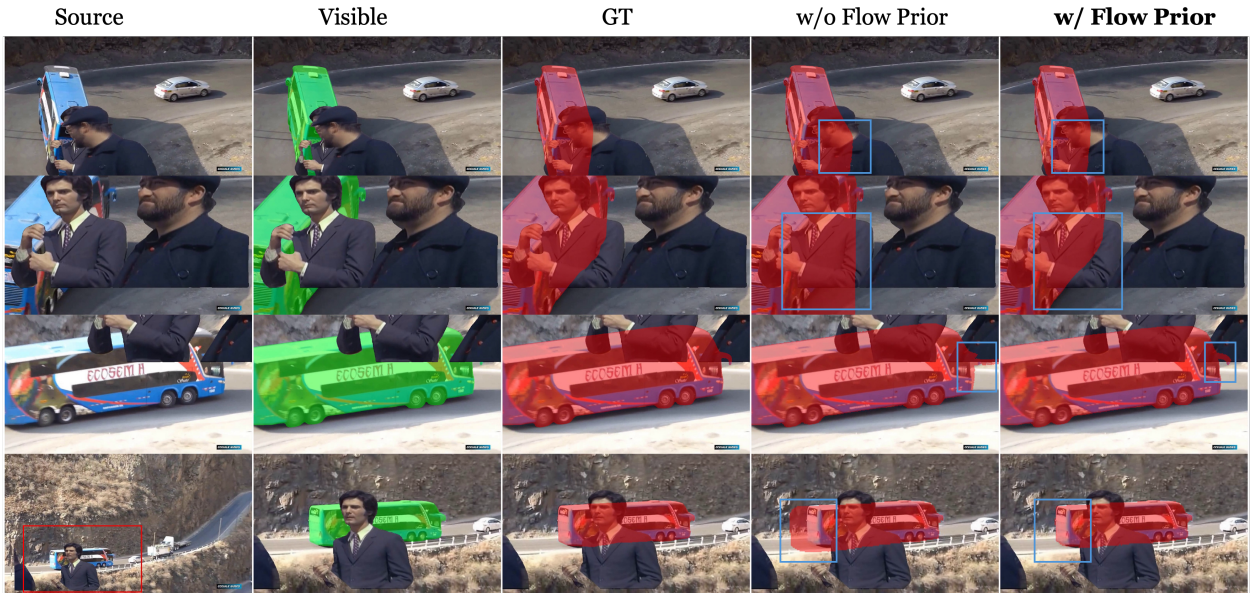

Figure 5: **Ablation of flow prior.** Amodal mask prediction with and without flow prior on LAVAT.

Quantitative Results. As shown in Table 2, REVEAL achieves state-of-the-art performance across all three benchmarks. On Kubric, REVEAL achieves substantial gains in texture completion (PSNR 29.92, SSIM 0.921, LPIPS 0.051) over the strongest baseline TACO (23.51 PSNR, 0.077 LPIPS), while also improving segmentation (mIoU 84.12) and reducing temporal artifacts (FVD 122.66). On TACO-OvO, REVEAL improves both completion and segmentation (PSNR 26.88, LPIPS 0.068, mIoU 95.12) with stronger temporal consistency (CLIP-F 0.988, FVD 213.77). On our long-horizon LAVAT-S, REVEAL achieves the best performance across all metrics, including CLIP-T (0.285), mIoU (90.29), and FVD (310.90), demonstrating the effectiveness of combining motion and semantic priors under occlusion. LAVAT-R contains 100 real-occlusion videos with human-annotated visible masks and text hints. Since amodal ground truth is unavailable, we evaluate via CLIP-T and a user study. As shown in Table 3, REVEAL achieves the best CLIP-T and highest user preference (71%), indicating more semantically aligned and perceptually preferred completions.

Qualitative Results. Figure 4 illustrates the qualitative advantage of REVEAL on LAVAT-R. In the first example, a man on a motorcycle is heavily occluded; TACO and Diff-VAS produce incomplete body and limb regions, whereas REVEAL reconstructs a more complete and coherent figure. In the second example, a rhinoceros is partially hidden behind boulders; competing methods generate incomplete shapes and artifacts, while REVEAL recovers the full body with consistent texture across frames. Qualitative comparisons against ground-truth amodal content on Kubric, TACO-OvO, and LAVAT-S are provided in the supplementary material (Figure IX, Figure VIII, and Figure VII, respectively). Video results are available on an anonymous webpage, see Appendix F. More qualitative results are available in the supplementary materials.

4.2.1 Ablation Study

Effect of flow mask priors guidance. Table 4 shows that the optical-flow mask prior consistently improves amodal segmentation across all occlusion levels, with gains that scale with occlusion severity. Under light occlusion (< 20%), the improvement is modest (+1.05 mIoU), as the visible mask already covers most of the

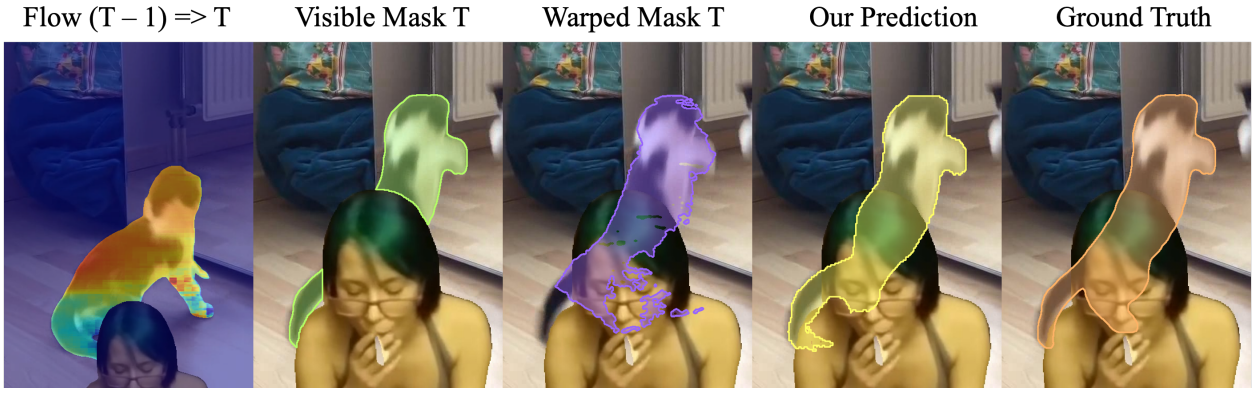

Figure 6: **Illustration of the flow prior under simultaneous deformation and occlusion.** The object deforms while being occluded. The visible mask at frame T (green) covers only the unoccluded portion. Warping the previous frame's mask to T via optical flow (purple) recovers regions that are currently hidden, providing a shape prior that captures the object's motion and deformation. Our predicted amodal mask (yellow) closely matches the ground truth (orange).

Table 5: **Flow-prior injection design ablation on LAVAT-S.** Added params is on top of the 397.66M-parameter Stage-1 trainable set; Latency is per 25-frame 128×256 fp16 step on a single A6000. mIoU (↑) is reported across occlusion difficulty.

Flow injection	**Added params**	**Latency (ms)**	< 20%	20–40%	40–65%	> 65%
w/o flow prior	0	86.9	91.88	87.97	85.75	87.79
Latent concat	11.5K	87.2 (+0.3)	92.20	88.50	86.40	88.10
MOFA-Adapter	735.2M	116.4 (+29.5)	92.80	90.30	87.95	88.80
Ours	0.635M	90.2 (+3.3)	**92.93**	**90.72**	**88.38**	**89.13**

object and leaves little ambiguity about its spatial extent. However, the benefit increases substantially under heavier occlusion: mIoU improves by +2.75 for 20–40% occlusion (87.97 → 90.72) and by +2.63 for 40–65% occlusion (85.75 → 88.38). Under extreme occlusion (> 65%), the flow prior still yields a meaningful gain of +1.34 mIoU (87.79 → 89.13). This trend aligns with our hypothesis: when more of the object is hidden, visible masks from previous frames—warped via optical flow—provide increasingly valuable evidence about the locations of occluded regions. Figure 5 confirms that flow priors help recover missing regions with cleaner boundaries, especially when the visible mask is fragmented. Figure 6 illustrates the case where the object simultaneously deforms and becomes occluded. The visible mask at frame T only covers the unoccluded portion, yet warping the previous frame's mask via optical flow recovers the hidden regions by capturing the object's motion and shape change, producing a prior that closely approximates the ground-truth amodal mask.

Flow-prior injection design. We compare three strategies for injecting the warped-mask flow signal into the Stage-1 U-Net, all using the same measured dense flow $\mathbf{F}_{k-1\to k}$ from DPFlow and identical training settings. (i) *Latent concat*: VAE-encode the warped previous-frame mask $\tilde{\mathbf{V}}_{k-1\to k}$ and append it as an extra input-channel block, so the input becomes $\mathbf{z}_t \oplus \mathcal{E}(\mathbf{V}) \oplus \mathcal{E}(\mathbf{D}) \oplus \mathcal{E}(\tilde{\mathbf{V}}_{k-1\to k})$. (ii) *MOFA-Adapter-style*, following MOFA-Video (Niu et al., 2024): a multi-scale Reference Encoder over $\mathbf{V}_{k-1}$, forward-warping under $\mathbf{F}_{k-1\to k}$, a Fusion Encoder mirroring the SVD encoder, and zero-conv residuals added into the frozen SVD decoder at every scale. (iii) *Ours*: a lightweight CNN g encodes the warped previous-frame mask (concatenated with $\mathbf{V}_k$) into a single per-frame token $\tilde{\mathbf{F}}_{\text{prior}}$ (Eq. 1), injected into the Stage-1 U-Net as a cross-attention KV stream (Eq. 2). Table 5 reports amodal mIoU on LAVAT-S by occlusion difficulty, with added trainable parameters and per-step latency. Cross-attention obtains the highest mIoU in every occlusion bucket at 0.635M added parameters and +3.3 ms latency. Latent concat yields only marginal gains

Table 6: **Text guidance ablation across occlusion difficulty on LAVAT.** We report completion quality for our Stage-2 texture model with and without text guidance.

	< 20%			20–40%			40–65%			> 65%		
Setting	SSIM↑	LPIPS↓	CLIP-T↑	SSIM↑	LPIPS↓	CLIP-T↑	SSIM↑	LPIPS↓	CLIP-T↑	SSIM↑	LPIPS↓	CLIP-T↑
w/o text guidance	0.982	0.026	0.281	0.945	0.087	0.282	0.945	0.087	0.279	0.942	0.090	0.269
w/ text guidance	0.987	0.021	0.286	0.960	0.067	0.285	0.959	0.071	0.284	0.957	0.072	0.273

Figure 7: **Qualitative ablation of text guidance.** Texture completion with and without text guidance on LAVAT. We also compare with VACE (Jiang et al., 2025) as an additional video inpainting baseline.

over the no-flow baseline despite the lowest cost, and MOFA-Adapter-style improves over latent concat but remains below cross-attention in every bucket while adding 735.2M parameters and +29.5 ms latency.

Figure 8: **Text cross-attention visualization.** Given the prompt "A man that is wearing a white shirt, a cap and a bag", we visualize the cross-attention maps for individual tokens. Each token attends to its corresponding semantic region, demonstrating that text guidance provides spatially grounded priors.

Table 7: **Flow prior ablation across occlusion duration on LAVAT-S.** Amodal mIoU (↑) over occluded frames, grouped by the number of consecutive frames the object has been occluded.

Occlusion duration (frames)	w/o flow prior	w/ flow prior	Δ
1–15	87.39	89.89	+2.50
16–40	88.03	90.61	+2.58
41–80	89.73	92.29	+2.56
81+	90.77	93.07	+2.30

Table 8: **End-to-end evaluation from plain-text input on LAVAT-S.** The first row upper-bounds Stage 2 with ground-truth amodal masks; the second row is our reported setting; the third row runs the full pipeline from the text prompt alone, with visible masks predicted by SAM3.

Visible mask	Amodal mask → Stage 2	mIoU↑	PSNR↑	SSIM↑	LPIPS↓	FVD↓
GT	GT amodal (Stage-2 upper bound)	—	29.55	0.972	0.058	276.4
GT	Stage-1 predicted (reported setting)	90.29	27.47	0.963	0.073	310.90
SAM3 (Carion et al., 2025)	Stage-1 predicted (end-to-end from text)	88.1	26.4	0.955	0.084	338.0

Effect of textual prior guidance. Table 6 demonstrates that text guidance improves completion quality across all occlusion regimes, with the benefit increasing as occlusion grows more severe. Under light occlusion ($< 20\%$), the model already reconstructs well from visible content alone, and text guidance provides only marginal gains (SSIM +0.005, LPIPS −0.005). As occlusion increases, the improvement grows substantially: at 20–40% occlusion, SSIM improves by +0.015 and LPIPS decreases by 0.020; at 40–65%, SSIM improves by +0.014 and LPIPS by 0.016. Under heavy occlusion ($> 65\%$), text guidance improves SSIM from 0.942 to 0.957 and reduces LPIPS from 0.090 to 0.072, yielding the largest absolute gains. This trend validates our core motivation: when visible content is insufficient to infer the complete appearance, text priors supply the missing semantic knowledge about the object's canonical shape and properties, leading to more faithful reconstructions. Qualitatively, Figure 7 shows that the unguided variant produces incomplete or washed-out regions. VACE (Jiang et al., 2025), as a video inpainting method, tends to fill occluded regions based on surrounding context rather than reconstructing the object as a whole, resulting in completions that blend with the background instead of preserving the object's own appearance. In contrast, our model with text guidance leverages semantic priors about the object category (e.g., "a man wearing a flannel shirt with dark color"), enabling coherent object reconstruction across frames. Figure 8 further visualizes the cross-attention maps for individual text tokens, showing that each token (e.g., "man", "shirt", "cap") attends to the corresponding semantic region, confirming that text guidance provides spatially grounded priors.

Effect of occlusion duration. Table 7 groups Stage-1 accuracy by the number of consecutive frames an object has been occluded. The flow prior yields a consistent gain of +2.3 to +2.6 mIoU across all buckets, from transient occlusions (1–15 frames) to long-lasting ones (81+ frames), showing that its benefit is invariant to occlusion duration.

End-to-end evaluation from plain-text input. Table 8 quantifies the error accumulation of the full pipeline on LAVAT-S. Feeding Stage 2 the Stage-1 predicted amodal mask instead of the ground-truth amodal mask (first vs. second row) costs ≈2.1 PSNR and +0.015 LPIPS. Replacing ground-truth visible masks with SAM3 (Carion et al., 2025) predictions obtained from the text prompt alone (second vs. third row), i.e., fully end-to-end from text, costs an acceptable gap ≈1.1 PSNR and 2.2 mIoU.

5 Conclusion & Discussion

We presented REVEAL, a diffusion-based framework for open-vocabulary video amodal completion that integrates complementary prior knowledge to address the limitations of existing methods. A user-provided text query serves a dual role: it drives open-vocabulary segmentation to obtain visible masks, and it acts as a semantic prior for texture reconstruction. In the first stage, optical flow serves as a motion prior, warping

visible masks from previous frames to propagate shape information into occluded regions, providing a strong shape prior particularly under heavy occlusion and when deformation and occlusion occur simultaneously. In the second stage, text guidance supplies object-class information through cross-attention for more faithful texture reconstruction, while providing a stable semantic reference that helps maintain visual coherence throughout the reconstruction. We also introduced LAVAT, a benchmark for text-guided video amodal completion featuring long video sequences paired with text descriptions, enabling evaluation of both temporal consistency and text-video alignment beyond the short clips used in prior work. Extensive experiments on Kubric, TACO-OvO, and LAVAT demonstrate that REVEAL achieves state-of-the-art performance across completion fidelity, segmentation accuracy, and temporal consistency.

Generalization to DiT-based video diffusion. Our two priors are architecture-agnostic conditioning signals. The latent-level conditioning $\mathbf{z}_t \oplus \mathcal{E}(\mathbf{V}) \oplus \mathcal{E}(\mathbf{D})$ ports to a video Diffusion Transformer (DiT) such as CogVideoX (Yang et al., 2024b) or WAN-Video (Wan et al., 2025) by widening the patch embedder's input projection. The flow prior $\tilde{\mathbf{F}}_{\text{prior}}$ maps onto a DiT cross-attention layer as an additional KV stream, and the Stage-2 text prior is already a CLIP-conditioned cross-attention native to DiTs. A DiT instantiation of REVEAL is left to future work.

Limitation. While text guidance improves completion quality by providing language priors, its effectiveness is primarily limited to category-level semantics. The text prompt can reliably guide the reconstruction of general object appearance and, to some extent, fine-grained properties such as texture and color. However, it does not yet consistently resolve ambiguities in more dynamic attributes, such as object pose or action (e.g., distinguishing between a running and walking motion). We provide a qualitative illustration of this failure mode in Appendix E. In addition, the flow prior requires the object to be visible in preceding frames: its gain is stable across occlusion durations (Table 7), but objects that remain heavily occluded from their first appearance provide no visible region to propagate, leaving such persistent occlusion outside the scope of motion-based priors.

References

Jiayang Ao, Yanbei Jiang, Qiuhong Ke, and Krista A Ehinger. Open-world amodal appearance completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6490–6499, 2025.

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*, pp. 5085–5092. IEEE, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23206–23217, 2023.

Kaihua Chen, Deva Ramanan, and Tarasha Khurana. Using diffusion priors for video amodal segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22890–22900, 2025.

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. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6144–6153, 2018.

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*, pp. 7346–7356, 2023.

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*, pp. 1272–1281, 2023.

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)*, pp. 569–585, 2018.

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*, pp. 1328–1336. IEEE, 2019.

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.

Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Coarse-to-fine amodal segmentation with shape prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1262–1271, 2023.

Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

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.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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*, 2022a.

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, 2022b.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13480–13492, 2023.

Cheng-Yen Hsieh, Tarasha Khurana, Achal Dave, and Deva Ramanan. Tracking any object amodally. *CoRR*, 2023.

Xun Huang, Zhengqi Li, Aleksander Holynski, Kartik Goyal, Noah Snavely, and Eli Park. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 127–135, 2015.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4019–4028, 2021.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ke Li and Jitendra Malik. Amodal instance segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 677–693. Springer, 2016.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17562–17571, 2022.

Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. *Advances in Neural Information Processing Systems*, 33:16246–16257, 2020.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.

Ruijie Lu, Yixin Chen, Yu Liu, Jiaxiang Tang, Junfeng Ni, Diwen Wan, Gang Zeng, and Siyuan Huang. Taco: Taming diffusion for in-the-wild video amodal completion. *arXiv preprint arXiv:2503.12049*, 2025.

Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17810–17820, 2025.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. MOFA-Video: Controllable image animation via generative motion field adaptations in frozen image-to-video diffusion model. In *European Conference on Computer Vision (ECCV)*, 2024.

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.

Matteo Poggi and Fabio Tosi. Flowseek: Optical flow made easier with depth foundation models and motion bases. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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.

Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2019.

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*, pp. 53–65. PMLR, 2020.

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*, pp. 8748–8763. PMLR, 2021.

N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9356–9366, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *In International Conference on Learning Representations*, 2021.

Minh Tran, Khoa Vo, Kashu Yamazaki, Arthur Fernandes, Michael Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. *In the 33rd British Machine Vision Conference*, 2022.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yihan Wang and Jia Deng. Waft: Warping-alone field transforms for optical flow. In *International Conference on Learning Representations*, 2026.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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*, pp. 7623–7633, 2023.

Rundi Wu, Ruoshi Liu, Carl Vondrick, and Changxi Zheng. Sin3DM: Learning a diffusion model from a single 3d textured shape. *The Twelfth International Conference on Learning Representations*, 2024.

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*, pp. 9099–9109, 2024.

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.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

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. *Advances in Neural Information Processing Systems*, 35:6278–6291, 2022.

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

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*, pp. 28003–28013, 2024.

Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3784–3792, 2020.

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*, pp. 586–595, 2018.

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*, pp. 7162–7172, 2024.

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