

Compositional Video Synthesis by Temporal Object-Centric Learning

Adil Kaan Akan, Yucel Yemez

We present a novel framework for compositional video synthesis that leverages temporally consistent object-centric representations, extending our previous work, SlotAdapt, from images to video. While existing object-centric approaches either lack generative capabilities entirely or treat video sequences holistically, thus neglecting explicit object-level structure, our approach explicitly captures temporal dynamics by learning pose invariant object-centric slots and conditioning them on pretrained diffusion models. This design enables high-quality, pixel-level video synthesis with superior temporal coherence, and offers intuitive compositional editing capabilities such as object insertion, deletion, or replacement, maintaining consistent object identities across frames. Extensive experiments demonstrate that our method sets new benchmarks in video generation quality and temporal consistency, outperforming previous object-centric generative methods. Although our segmentation performance closely matches state-of-the-art methods, our approach uniquely integrates this capability with robust generative performance, significantly advancing interactive and controllable video generation and opening new possibilities for advanced content creation, semantic editing, and dynamic scene understanding.



Index Terms—Object-centric learning, compositional video generation and editing, unsupervised video object segmentation, slot diffusion, invariant slot attention

1 INTRODUCTION

The real world is inherently compositional, made up of distinct entities that can be flexibly combined and reconfigured into richer structures. This property underlies core cognitive abilities such as abstraction, causal reasoning, and systematic generalization [19, 35, 5]. However, beyond static structure, real-world environments are also deeply dynamic: objects move, interact, appear, and disappear over time. Humans naturally parse these dynamics into persistent, interacting entities, forming temporally coherent mental models that support understanding and prediction [59, 60]. Modeling such temporal compositionality remains a major challenge for artificial systems. While object-centric learning has shown promise in uncovering latent structure in images [22], extending these ideas to video requires capturing not only spatial grouping, but also object continuity, interaction, and transformation across time. Although recent text-to-video models have achieved impressive synthesis quality, they generally lack compositional structure and offer limited control over object-level content. Bridging this gap is essential for building video models that can reason,

generalize, and interact with dynamic environments in a human-like, compositional manner.

Recent progress in object-centric learning (OCL) has been especially notable in static image domains, with models such as SlotDiffusion [68], Latent Slot Diffusion (LSD) [27], and our prior work SlotAdapt [3] achieving strong results in unsupervised object discovery, segmentation, and image generation. These methods decompose scenes into discrete, interpretable object representations (“slots”) [35, 52], enabling structured understanding and high-fidelity synthesis. Importantly, they offer a pathway toward models that can generalize compositionally by reasoning over object-level primitives.

In our previous work, SlotAdapt [3], we introduced a framework that leverages pretrained diffusion models conditioned with slot-based representations via adapter layers. It outperformed earlier slot-based approaches in both segmentation and image generation and enabled compositional editing of real-world images—an ability that prior models lacked. However, generative modeling from object-centric representations in video remains largely unexplored.

The broader OCL field has gradually progressed from synthetic datasets [28, 29] to real-world images [18, 37] and videos [45, 70, 71, 36, 43], typically within an autoencoding framework [21, 39]. These models aim to uncover object structure by reconstructing input frames using architectural or data-driven priors, often guided by static cues like color, shape, or pretrained features.

Extending object-centric generative modeling to the video domain introduces unique challenges. Models must capture temporal continuity, dynamic object transforma-

• Adil Kaan Akan and Yucel Yemez are with Department of Computer Engineering, Koc University. Yucel Yemez is with KUIS AI Center, Koc University. E-mail: kakan20@ku.edu.tr, yyemez@ku.edu.tr

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

tions, and multi-object interactions across frames. Most existing approaches focus on segmentation or tracking and enforce temporal consistency through auxiliary signals such as optical flow [30] or depth maps [15], or by incorporating architectural biases that promote slot stability [56, 67] such as slot-aligned cross-frame attention and slot-specific recurrent updates. While such cues can aid learning, they introduce substantial computational overhead and are fragile in the presence of motion blur, deformation, or occlusion. Recent works in unsupervised temporal OCL, such as SOLV [4], TC-Slot [41], and others [14, 47], explore learning temporally consistent slots using pretrained vision encoders (e.g., DINO [10, 53] or CLIP [48]). While effective in producing stable representations, these methods lack generative capabilities and compositional control, relying on feature-level decoding instead of pixel-level synthesis. As a result, they cannot support interactive editing or video generation.

Meanwhile, diffusion-based video generation models [25, 54, 9] have demonstrated high-quality synthesis using large-scale text-video datasets. However, these models treat scenes as monolithic visual fields and do not incorporate object-centric structure. This limits compositional controllability, semantic disentanglement, and consistent object identity over time. Some recent efforts [64, 7, 73] introduce control mechanisms through keyframes, masks, or motion guidance, but they remain orthogonal to our approach and do not utilize slot-level object representations.

In this work, we propose a fully self-supervised framework for generative video modeling that combines object-centric representation learning with high-quality synthesis capabilities. Building on SlotAdapt [3], we extend object-centric generation into the temporal domain by learning temporally consistent slots that encode object identity, motion, and interactions—without relying on external signals like optical flow or depth. By conditioning a pretrained diffusion model on these slot-based temporal features, we achieve video reconstructions that are both temporally coherent and semantically grounded through pixel-level diffusion synthesis.

A key methodological distinction lies in our model’s ability to discover temporal structure directly from raw video data. Rather than relying on architectural constraints such as STEVE’s deterministic slot transitions [56] or SlotFormer’s slot-aligned cross-frame attention [67] or hand-designed temporal signals like optical flow guidance in G-SWM and STOVE [38, 32], our approach learns dynamic object relationships through self-supervised conditioning of a pretrained diffusion model with slot-based temporal features. This design enables object-centric video generation with compositional editing control, capabilities absent in existing paradigms.

Unlike conventional text-to-video models [25, 54] that operate holistically without object-level structure, or feature-decoding temporal OCL methods [4, 71] that lack generative capabilities, our slot-based method supports flexible compositional control, allowing users to insert, remove, or modify individual objects while maintaining temporal coherence and scene consistency. Empirically, our model achieves competitive segmentation performance compared to existing unsupervised methods such as SOLV [4] or OCLR [69], while also enabling full-resolution video syn-

thesis. Compared to image-based OCL models including LSD [27], SlotDiffusion [68], and SlotAdapt [3], our approach establishes new benchmarks for video synthesis with substantial improvements in temporal consistency and visual fidelity. To the best of our knowledge, our method is the first to combine object-centric representation learning with high-quality video generation on real-world videos, enabling compositional editing where individual objects can be inserted, removed, or modified while maintaining temporal coherence, in a self-supervised framework.

2 RELATED WORK

Unsupervised Object-Centric Learning (OCL). Unsupervised object-centric learning aims to decompose visual scenes into discrete and semantically meaningful representations, typically referred to as “slots”, without explicit supervision. Early works such as Attend-Infer-Repeat (AIR) [16] and SQAIR [31] used iterative inference and variational decoders but were limited to simple, low-complexity settings. Slot Attention [39] introduced a permutation-invariant attention mechanism that enabled effective object discovery and segmentation on synthetic datasets. This approach has since been extended with autoregressive transformers (e.g., SLATE [55], discrete tokenization [56], and self-supervised objectives in feature space (e.g., DINOSAUR [53]). More recently, hierarchical approaches have been explored, such as COCA-Net [33], which introduces a hierarchical clustering strategy with spatial broadcast decoding to achieve superior segmentation performance on synthetic datasets.

Several methods have explored temporal extensions of object-centric modeling by encouraging slot consistency across frames. SAVi [30] and SAVi++ [15] incorporate predictor/corrector architecture which relies on auxiliary signals such as optical flow and depth to guide temporal slot alignment. However, such additional cues are prone to failure under deformation, occlusion, or motion blur. To address these limitations, self-supervised alternatives have been proposed: TC-Slot [41] employs contrastive learning to enforce cross-frame slot consistency; Betrayed-by-Attention [14] introduces a combination of hierarchical clustering and consistency objective to stabilize slot attention; RIV [47] reconsiders image-to-video transfer from an object-centric lens; and SOLV [4] applies Invariant Slot Attention [8] to cluster DINO features across time for coherent unsupervised slot assignments. Early work in future prediction and scene understanding [1, 2] also highlighted the relevance of object-centric temporal modeling, motivating subsequent developments in this direction.

Despite these advances in segmentation and tracking, most temporal OCL methods rely on feature-level decoding and thus lack pixel-level generative capabilities—crucial for photorealistic synthesis and fine-grained control. Moreover, they do not support structured or compositional manipulation of real-world content. In contrast, our method significantly advances this direction by enabling pixel-level, temporally consistent video generation and slot-based compositional editing in complex, real-world scenarios.

Diffusion Models for Image Generation. Diffusion models [58, 24] have rapidly become the state-of-the-art for high-

fidelity image generation, with models like ADM [13], Imagen [51], and DALLÉ-2 [49] showcasing controllable, text-conditioned synthesis. Latent Diffusion Models (LDMs) [50] reduce the computational cost by operating in compressed latent space and offer flexible conditioning through cross-attention. Several recent works have aimed to scale or improve the underlying architecture [34, 17, 46, 11]. Recent extensions like T2I-Adapters [42] allow lightweight adaptation of pretrained diffusion models to new tasks without retraining the core model.

Object-Centric Diffusion Models. Several recent works combine object-centric learning with diffusion-based generation to improve compositionality and controllability in images. LSD [27], SlotDiffusion [68], GLASS [57] and our prior work SlotAdapt [3] use slot-based representations to condition diffusion decoders, enabling unsupervised object discovery and structured image generation. However, LSD and GLASS [57] rely on pretrained diffusion models, and often suffer from a mismatch between learned slots and the pretrained attention layers, leading to degraded generative quality. SlotDiffusion addresses this by training the diffusion model from scratch, but this approach requires extensive compute and lacks generalization.

In our previous work, SlotAdapt [3], we introduce adapter layers specifically designed to better align slot semantics with pretrained diffusion priors, achieving superior segmentation and image generation performance compared to prior slot-based methods. Furthermore, SlotAdapt was the first to successfully demonstrate compositional generation and editing capabilities on challenging real-world image datasets. Despite these advancements, existing object-centric diffusion models—including SlotAdapt—primarily focus on static images, leaving the challenge of modeling temporal object dynamics largely unaddressed. Recognizing this crucial gap, our current work explicitly extends SlotAdapt’s foundational ideas into the temporal domain, enabling coherent object-centric video synthesis and editing. We significantly advance beyond previous image-centric methods by incorporating mechanisms to model object continuity and interaction over time.

Video Generation with Diffusion Models. Recent advances in video synthesis have been driven by diffusion models capable of producing high-resolution, temporally consistent outputs. Ho et al. [26] introduced Video Diffusion Models (VDMs), extending denoising diffusion to sequential data. Follow-up models such as Imagen Video [25], Make-A-Video [54], Phenaki [63], and Align Your Latents [9] further advanced text-to-video generation using large-scale paired datasets. Despite their impressive results, these models operate on holistic, entangled scene representations and lack explicit object-level structure. This limits their ability to support semantic disentanglement, compositional reasoning, or precise object control.

While there have been early steps toward compositionality in video—e.g., Tune-A-Video [66], VideoComposer [64], and ControlVideo [73]—these methods rely on fine-tuning or auxiliary conditioning (e.g., masks, keyframes, or motion) rather than on disentangled object representations. As such, they lack an object-centric inductive bias and cannot perform structured manipulations such as adding, removing, or replacing individual entities in a scene.

In contrast, our work addresses this fundamental limitation by learning object-centric temporal representations that enable direct generative control. Unlike existing approaches that operate post-hoc on pretrained models [66, 64, 73, 42], we develop an end-to-end framework that jointly learns object discovery and temporally consistent synthesis from raw video data.

3 PRELIMINARIES

3.1 Slot Attention

Slot Attention [39] provides a framework to decompose visual scenes into discrete, interpretable components termed *slots*. Slots are initialized randomly and iteratively refined to represent distinct objects or entities within the input data through an attention mechanism. Formally, the slot update rule can be described as:

$$\mathbf{U}^{(m)} = \text{Attention}\left(q(\mathbf{S}^{(m)}), k(\mathbf{F}), v(\mathbf{F})\right), \quad (1)$$

$$\mathbf{S}^{(m+1)} = \text{GRU}(\mathbf{S}^{(m)}, \mathbf{U}^{(m)}), \quad (2)$$

where q , k , and v represent learnable linear transformations corresponding to queries, keys, and values, respectively; \mathbf{F} denotes extracted image features; \mathbf{S} are the slot vectors; \mathbf{U} represents the update generated by the attention operation; m is the iteration index. Slots compete for pixels through attention over the slot dimension, encouraging each slot to bind to a distinct region or object in the scene. The effectiveness of Slot Attention lies in its capacity to disentangle and encode complex scenes into structured representations without supervision. This is achieved through the ability of the mechanism to assign representational capacity where needed.

3.2 Diffusion Models

Diffusion models [58, 24] have recently emerged as powerful generative frameworks, producing high-quality samples by modeling the reverse process of a progressive noise addition. Given an input image \mathbf{X} , the model learns to reverse a noisy image \mathbf{X}_τ at timestep τ^1 back to the original data distribution.

The underlying generative process consists of a series of reverse diffusion steps that transform samples from a noise distribution to the data distribution $p(\mathbf{X})$, where each step is defined by conditional probabilities $p(\mathbf{X}_{\tau-1}|\mathbf{X}_\tau)$. During training, a loss function \mathcal{L} is optimized that penalizes the expected prediction error across random timesteps τ , effectively teaching the model to denoise progressively corrupted inputs and reconstruct the original data distribution via:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X} \sim p(\mathbf{X}), \epsilon_\tau \sim \mathcal{N}(0,1), \tau \sim \mathcal{U}(1,T)} [\|\epsilon_\tau - \epsilon_\theta(\mathbf{X}_\tau, \tau, y)\|_2^2], \quad (3)$$

where ϵ_θ denotes the noise prediction network parameterized by $\boldsymbol{\theta}$, and y represents optional conditioning information. For high-resolution images, Latent Diffusion Models (LDM) [50] are proposed to perform the training and sampling in a low-dimensional latent space, obtained through a pretrained variational autoencoder (VAE) [50]. This approach improves computational efficiency while preserving generative fidelity by decoupling image compression and synthesis.

1. We use the symbol τ for diffusion timesteps to distinguish from t which denotes video frame index throughout this paper.

3.3 SlotAdapt

In our prior work, SlotAdapt [3], we introduced a robust methodology that integrates the strengths of Slot Attention and pretrained diffusion models to achieve effective object-centric image generation and compositional editing. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, SlotAdapt first employs a visual backbone, typically based on pretrained vision transformers like DINO [10, 44], to extract a compact set of visual features represented by $\mathbf{F} \in \mathbb{R}^{h \cdot w \times d}$, where h, w indicate reduced spatial dimensions and d is the feature dimension. Slot Attention is then applied to these features to dynamically allocate object-centric representations into discrete slots $\mathbf{S} \in \mathbb{R}^{K \times D_{\text{slot}}}$, with each slot ideally corresponding to a separate object or distinct entity in the scene.

These learned slots condition a pretrained Stable Diffusion decoder, which we augment with specialized adapter layers. Specifically, these adapter layers, implemented as additional cross-attention modules inserted into each down-sampling and up-sampling block of the pretrained U-Net architecture, enable explicit conditioning on slot representations. This design significantly differs from prior approaches such as SlotDiffusion [68] and LSD [27], which condition on slots via cross-attention layers originally trained for text embeddings—leading to potential misalignment between slot semantics and the pretrained diffusion attention pathways. By separating the adapter conditioning from the text-based conditioning modules of the pretrained diffusion model, we allow slots to focus exclusively on object-level semantics, independent from the original textual embedding space.

To further enhance the conditioning mechanism, SlotAdapt introduces a dedicated register token, \mathbf{r} , computed by mean pooling the slot representations or alternatively, the visual backbone features, following [12]. This register token captures overall contextual scene information and is conditioned via the original text-based cross-attention layers of the diffusion model, enabling slots to remain focused on specific objects without being diluted by global scene context.

During training, SlotAdapt utilizes a reconstruction-based loss framed as a noise prediction objective within the diffusion framework. Formally, given a noisy latent image representation \mathbf{X}_τ obtained at diffusion timestep τ , SlotAdapt aims to predict the noise ϵ_τ using slot \mathbf{S} and register \mathbf{r} conditioning:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{X} \sim p(\mathbf{X}), \epsilon_\tau \sim \mathcal{N}(0, \mathbf{I}), \tau \sim \mathcal{U}(1, T)} \left[\|\epsilon_\tau - \epsilon_\theta(\mathbf{X}_\tau, \tau, \mathbf{S}, \mathbf{r})\|_2^2 \right], \quad (4)$$

where ϵ_θ denotes the diffusion-based noise prediction model parameterized by θ .

Importantly, SlotAdapt freezes the pretrained diffusion model parameters throughout training, updating only the adapter layers and the slot attention mechanism. This strategy allows the method to efficiently leverage the powerful generative prior encoded in pretrained diffusion models, significantly enhancing generative performance and training efficiency. Additionally, SlotAdapt introduces a guidance loss that enforces alignment between slot attention masks and the diffusion model’s cross-attention maps, leveraging the prior knowledge residing in the frozen diffusion layers to improve slot-object correspondence during training.

SlotAdapt achieves state-of-the-art results on object-centric image generation and compositional editing, particularly on complex real-world datasets. Its robust methodology, which combines structured object representations with pretrained diffusion, provides a strong foundation for this work, where we extend these ideas into the temporal domain for object-centric video generation and editing.

4 METHODOLOGY

Our proposed framework extends SlotAdapt [3] to generate high-quality, temporally coherent videos from object-centric representations. It comprises two core components: a temporal object-centric encoder that captures dynamics and interactions across video frames, and a slot-conditioned diffusion decoder that synthesizes photorealistic frames. The entire architecture is trained in a fully self-supervised manner, without reliance on auxiliary cues such as optical flow or depth.

4.1 Object-Centric Temporal Encoding

Given an input video composed of L frames, our model first extracts visual features from each frame using a frozen DINOv2 backbone [44]. Specifically, each frame at timestep t is transformed into patch-level features $\mathbf{F}_t \in \mathbb{R}^{N \times d}$, where each patch corresponds to a distinct spatial region of the frame, N is the number of total patches and d is the feature dimension.

To achieve temporally consistent object-centric representations, we utilize Invariant Slot Attention (ISA) [8]. While original ISA enforces spatial invariance by decoupling object identity from spatial information, our approach retains ISA’s slot attention mechanism but replaces the spatial broadcast decoder with a diffusion decoder that is directly conditioned on pose-invariant slots, with register tokens providing the necessary spatial context for coherent generation (discussed in Section 4.3).

Formally, ISA decomposes each frame into K slot vectors, each slot vector $\mathbf{z}_t^j \in \mathbb{R}^{D_{\text{slot}}}$ attending to frame features as follows:

$$\mathbf{A}_t := \text{softmax}_{j=1, \dots, K}(\mathbf{M}_t) \in \mathbb{R}^{K \times N}, \quad (5)$$

$$\mathbf{m}_t^j := \frac{1}{\sqrt{d}} p \left(k(\mathbf{F}_t) + g(\mathbf{G}_{\text{rel}, t}^j) \right) q(\mathbf{z}_t^j) \in \mathbb{R}^N \quad (6)$$

where $q : \mathbb{R}^{D_{\text{slot}}} \rightarrow \mathbb{R}^{D_{\text{slot}}}$, $k : \mathbb{R}^d \rightarrow \mathbb{R}^{D_{\text{slot}}}$, $p : \mathbb{R}^{D_{\text{slot}}} \rightarrow \mathbb{R}^{D_{\text{slot}}}$, and $g : \mathbb{R}^2 \rightarrow \mathbb{R}^{D_{\text{slot}}}$ are learnable linear projections applied to each patch and slot vector independently and $\mathbf{G}_{\text{rel}, t}^j \in \mathbb{R}^{N \times 2}$ encodes the relative spatial position of each patch with respect to slot j using a learnable scale and shift transformation. The unnormalized attention scores $\mathbf{m}_t^j \in \mathbb{R}^N$ represent the j -th row of the matrix $\mathbf{M}_t \in \mathbb{R}^{K \times N}$, computed for all K slots. The softmax is then applied column-wise (over slots) to obtain the attention matrix $\mathbf{A}_t \in \mathbb{R}^{K \times N}$. The attention matrix is computed per frame, where each row represents the contribution of each patch to the update of slot \mathbf{z}_t^j at time t . Please refer to the supplementary material for details on how we adapted ISA to our case.

These slot vectors are iteratively refined using a Gated Recurrent Unit (GRU) and residual Multi-Layer Perceptron (MLP), following the Slot Attention mechanism [39], to reinforce their object binding consistency across frames. The output per frame is a set of K slots, denoted by

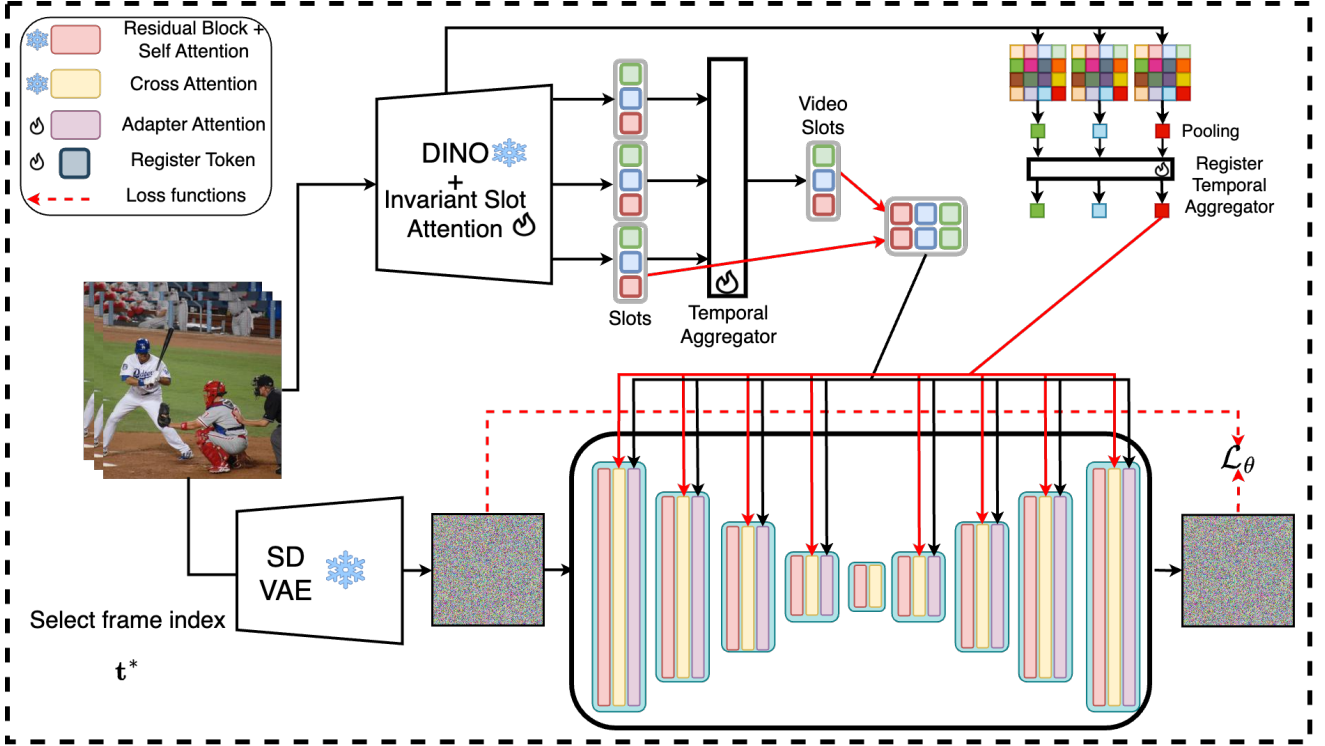


Fig. 1. **Architecture Block Diagram** We extract object-centric and temporally consistent information from input video frames using a visual backbone composed of DINOv2 and Invariant Slot Attention (ISA). The ISA mechanism generates slots for each frame, which are then aggregated temporally using a Transformer-based temporal aggregator to produce enriched, temporally-aware video slots. Concurrently, global context information is summarized by average pooling frame-level features and further processed through a separate temporal aggregator to produce global scene tokens. A pretrained Stable Diffusion Variational Autoencoder (VAE) encodes a randomly selected frame into latent space, and Gaussian noise is subsequently added. The diffusion model is explicitly conditioned on both the temporally aggregated video slots and the slots from the selected frame (shown here as the last frame for visualization purposes, though in practice this could be any frame) via additional adapter attention layers, and on the global scene token using the diffusion model's native cross-attention layers. During training, the model learns to predict the added noise, with the diffusion loss (\mathcal{L}_θ) measuring the prediction error. This training strategy ensures temporally coherent, object-centric video synthesis and intuitive compositional editing capabilities across video frames.

$\mathbf{S}_t = \{\mathbf{z}_t^1, \dots, \mathbf{z}_t^K\}$, accompanied by corresponding soft attention masks \mathbf{A}_t .

Temporal Context Aggregation. To capture broader temporal context, we concatenate slot sequences from all frames within the video segment and pass them through a Transformer encoder augmented with learnable temporal positional embeddings:

$$\tilde{\mathbf{S}}_{1:L} = \text{Transformer}(\mathbf{S}_{1:L}), \quad \mathbf{S}_{1:L} = \text{concat}(\mathbf{S}_1, \dots, \mathbf{S}_L). \quad (7)$$

The resulting temporally aggregated slots $\tilde{\mathbf{S}}_{1:T}$ are reshaped and concatenated back with original frame-wise slots, creating augmented slots for each frame:

$$\tilde{\mathbf{S}}_t^+ = \text{concat}(\mathbf{S}_t, \tilde{\mathbf{S}}_t). \quad (8)$$

Additionally, we compute global scene-level context vectors by average-pooling the DINO features of each frame. These pooled vectors undergo further temporal encoding via another Transformer encoder to yield temporally-aware global tokens $\tilde{\mathbf{r}}_t \in \mathbb{R}^d$:

$$\mathbf{r}_t = \frac{1}{N} \sum_{i=1}^N \mathbf{F}_{t,i}, \quad \tilde{\mathbf{r}}_t = \text{Transformer}(\mathbf{r}_1, \dots, \mathbf{r}_L)_t. \quad (9)$$

These global tokens summarize high-level dynamics and semantic context, complementing the slot-based object representations.

During training, we adopt a 1-frame training strategy: we randomly select a single frame index t^* at each iteration and retain only its temporally enriched slot set $\tilde{\mathbf{S}}_{t^*}^+$ and

global context token $\tilde{\mathbf{r}}_{t^*}$ for decoding. This 1-frame training allows efficient supervision using a pretrained image-based diffusion model.

4.2 Slot-Conditioned Diffusion Decoding

We decode the selected video frame using a pretrained Stable Diffusion model [50]. The chosen frame \mathbf{V}_{t^*} is encoded into a latent representation $\mathbf{X}_{t^*} \in \mathbb{R}^{h \times w \times c}$ via the VAE encoder of the diffusion model. The latent representation is then perturbed by Gaussian noise according to the diffusion schedule, resulting in a noisy latent \mathbf{X}_τ at timestep τ .

To condition the diffusion process explicitly on object-level semantics, we inject lightweight adapter-based cross-attention layers into each residual block of the U-Net decoder, following the SlotAdapt architecture [3]. Adapter layers are used to condition the diffusion U-Net with the augmented slots, $\tilde{\mathbf{S}}_{t^*}^+$. Concurrently, the native cross-attention layers of the U-Net, originally designed for textual embeddings, are used to condition the global scene token $\tilde{\mathbf{r}}_{t^*}$, which effectively summarizes global contextual information.

The diffusion training objective aims to predict the added noise:

$$\mathcal{L}_{\text{diff}} = \left\| \epsilon - \epsilon_\theta \left(\mathbf{X}_\tau, \tau, \tilde{\mathbf{S}}_{t^*}^+, \tilde{\mathbf{r}}_{t^*} \right) \right\|^2, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (10)$$

Throughout training, we freeze the pretrained diffusion model parameters and optimize only the temporal slot encoder and the adapter layers. This ensures that the model

benefits from the robust generative prior inherent in the pretrained diffusion backbone without requiring large-scale retraining.

Unlike SlotAdapt, we omit auxiliary attention-guidance losses. Since we employ 1-frame training (randomly selecting one frame per iteration for efficient decoding), applying guidance loss would only align slot attention masks with diffusion cross-attention masks for the selected frame, while other frames in the temporal window receive no such alignment signal. This inconsistent gradient flow across frames disrupts temporal coherence. Instead, we directly utilize encoder-derived attention masks without additional processing or merging, ensuring simplicity and stability.

Inference. At test time, we apply a sliding-window strategy, decoding the central frame within each overlapping window independently, following SOLV [4]. We align slot identities across frames using Hungarian matching based on slot similarity. This design supports temporally coherent video synthesis and provides intuitive compositional editing capabilities: objects can be explicitly manipulated across frames by modifying corresponding slots directly. In contrast to prior methods, which were limited to synthetic videos or static images, our approach successfully handles real-world dynamic video scenes, effectively bridging segmentation accuracy and high-quality generative performance.

4.3 Invariant Slot Attention Adaptation for Diffusion Conditioning

The effectiveness of our object-centric video generation framework relies on the assumption that the slot representations are temporally consistent and invariant to pose changes induced by motion. To achieve this, we employ ISA on the encoder side, which estimates slot-specific pose parameters and incorporates them into relative position encodings during slot formation. In the original ISA architecture, these pose estimates are also used during decoding via a spatial broadcast mechanism, enabling spatially accurate reconstructions from pose-invariant slots.

In our model, while the slot attention computation remains identical to that of the original ISA, we replace the spatial broadcast decoder with a pretrained diffusion decoder to enable high-quality generative modeling. This decoder is directly conditioned on the pose-invariant slots, which by design lack explicit spatial information. To compensate for this, we leverage the register tokens (introduced in Sections 3.3 and 4.1), which provide global pose and background context to the diffusion decoder. The slots and register tokens together maintain a disentangled representation of object identity and spatial attributes, enabling spatially coherent video generation within a diffusion-based framework.

To empirically verify the role of register tokens, we generate images with and without register tokens (Fig. 2) using a model trained with both components. When register tokens are omitted (replaced with zero vectors), the generated objects appear in incorrect positions, with incorrect scales and, in some cases, altered orientations compared to the ground truth. The backgrounds also deviate significantly from the original scenes. In contrast, when register tokens are included, the generated object positions, scales, orientations, and backgrounds closely match the ground truth.

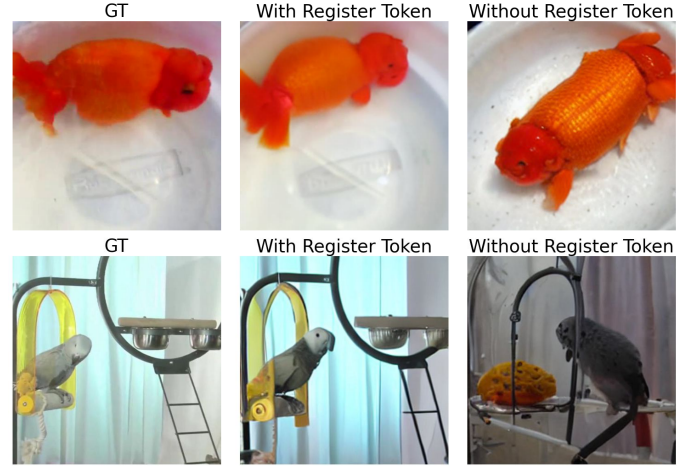


Fig. 2. **Pose Invariance in Diffusion Conditioning.** Comparison of video frame generation with and without register tokens on YTVIS dataset. **Without register tokens** (middle), objects appear in incorrect positions and backgrounds deviate from ground truth. **With register tokens** (right), generations accurately match ground truth (left), confirming that register tokens handle pose information while slots maintain object identity focus. Full temporal sequences are provided in the Supplementary Material.

These results demonstrate that the register tokens effectively capture the pose and spatial context information that would otherwise be handled by the spatial broadcast decoder in the original ISA architecture. Please refer to the Supplementary Material for additional results on the role of register tokens in capturing spatial context.

5 EXPERIMENTS

We comprehensively evaluate our proposed framework against state-of-the-art object-centric learning methods. We focus on two core tasks: unsupervised video object segmentation and temporally consistent video generation. Experiments are conducted on two real-world datasets using widely adopted metrics to ensure rigorous and meaningful evaluation.

5.1 Datasets

We evaluate our approach on two widely used real-world video datasets: DAVIS17 [45] and YouTube-VIS 2019 (YTVIS19) [71].

DAVIS17 is a benchmark specifically tailored for video object segmentation. It contains short, high-quality video sequences annotated with precise ground-truth masks, requiring temporal consistency.

YTVIS19 consists of diverse video sequences with complex scenes and significant variation in object appearance, pose, and background. Following prior work [4], we evaluate our model on a subset comprising 300 videos selected from the original training set of 2,883 high-resolution videos, as YTVIS19 lacks an official validation or test set with provided ground-truth masks.

Together, these datasets provide a challenging and realistic evaluation environment for both segmentation and generation tasks.

5.2 Implementation Details

We follow the implementation practices from our prior work, SlotAdapt [3], adapting and extending them to

TABLE 1

Ablation study on YTVIS dataset. We systematically evaluate the contribution of key components in our unified framework by comparing against our full model configuration. The full model uses invariant slot attention, DINO register tokens, register aggregator, and 1-frame training. We analyze the impact of removing individual components and varying training strategies.

Method Configuration	mIoU	FG-ARI
Full Model	40.57	22.40
<i>Component Ablations</i>		
w/o Register Aggregator	39.22	20.49
w/ Slot Avg Register Tokens	36.87	18.06
w/ Standard Slot Attention	27.09	11.06
<i>Training Strategy Ablations</i>		
5-frame Training	40.67	22.00
5-frame Training + Guidance	41.02	22.69

model temporal dynamics. Specifically, we use a frozen DINOv2 [44] ViT-B/14 as the visual backbone to extract frame-level features. Invariant Slot Attention (ISA) is applied per frame with shared initialization across time. A transformer-based temporal aggregator enriches the slots with temporal context, using a temporal window of $L = 5$ frames (2 past, 1 present, 2 future) following previous work [4].

For decoding, we use Stable Diffusion v1.5 [50], with adapters inserted as in SlotAdapt [3]. During training, we keep the Stable Diffusion model parameters fixed and optimize only the ISA, the temporal transformers, and the adapter layers. We train all models for 350K iterations on YTVIS19, then fine-tune for 50K iterations on DAVIS17. All experiments are conducted on 2× NVIDIA A40 GPUs with 48GB memory each.

5.3 Baselines

For segmentation, we compare against SOLV [4], a recent state-of-the-art method in unsupervised temporal object-centric learning.

For the video generation task, no prior method directly addresses object-centric video generation from unsupervised representations, as existing object-centric approaches use feature decoders such as those proposed by DINO-SATUR [53]. Thus, we benchmark against existing object-centric image generation models: SlotDiffusion [68], Latent Slot Diffusion (LSD) [27], and our previously developed SlotAdapt [3] method, by training them on video frames individually. This provides a rigorous baseline, highlighting our method’s unique capability of generating coherent videos directly from object-centric representations that maintain temporal coherence.

5.4 Evaluation Metrics

Segmentation: We employ two complementary evaluation metrics for comprehensive segmentation assessment. We use the Foreground Adjusted Rand Index (FG-ARI) to measure the quality of clustering foreground pixels into distinct object segments. Following prior work [4, 29, 6], we calculate per-frame FG-ARI and report the mean across all frames for consistency with existing approaches.

Additionally, we utilize mean Intersection-over-Union (mIoU) focusing on foreground objects, which is widely accepted in segmentation literature [53]. To ensure temporal consistency between frames, we apply Hungarian matching

TABLE 2

Unsupervised video object segmentation on real-world datasets. We compare our method with state-of-the-art approaches on YTVIS and DAVIS datasets. For fair comparison with our encoder-based approach, we include SOLV-E (encoder attention masks) and SOLV-E + M (encoder attention masks with merging), alongside the full SOLV method which uses decoder-generated masks. Our approach demonstrates strong performance across both clustering-based (FG-ARI) and overlap-based (mIoU) evaluation metrics.

Method	YTVIS		DAVIS	
	mIoU	FG-ARI	mIoU	FG-ARI
LSD [27]	29.55	14.07	29.55	14.35
SlotDiffusion [68]	38.33	15.70	31.27	12.34
SlotAdapt [3]	36.51	20.32	29.95	16.28
SOLV-E	32.91	19.30	31.23	18.89
SOLV-E + M	36.91	21.34	33.12	20.40
SOLV [4] ²	42.01	21.55	36.62	20.98
Ours	40.57	22.40	34.93	21.60

between predicted and ground-truth masks following standard practice [45].

Generation: We evaluate video generation quality through a number of complementary metrics that capture different aspects of visual fidelity and perceptual quality. We employ Peak Signal-to-Noise Ratio (PSNR) to quantify pixel-wise reconstruction accuracy between generated and ground-truth frames. To assess perceptual similarity, we utilize the Structural Similarity Index (SSIM) [65], which evaluates structural information preservation, including luminance and contrast patterns. For deeper perceptual evaluation, we incorporate Learned Perceptual Image Patch Similarity (LPIPS) [72], which leverages deep features to measure perceptual differences that correlate with human judgments. To evaluate distributional quality and realism, we employ Fréchet Inception Distance (FID) [23], which measures feature distribution distances between real and generated images. Additionally, we utilize Fréchet Video Distance (FVD) [61] to assess temporal consistency and motion quality in generated video sequences. This multi-perspective evaluation ensures a robust understanding of generation quality in both spatial and temporal dimensions.

5.5 Quantitative Results

We first analyze the contribution of individual components through ablation studies, then compare our unified framework against specialized baselines for both segmentation and generation tasks.

Ablation Studies. We systematically investigate our framework’s key components and training strategies (Table 1). Removal of the register aggregator (the version where only the corresponding frame’s DINO tokens are used without any transformer encoder for temporal aggregation, \mathbf{r}_t in Eq. 9) or replacing register tokens with slot averaging (the augmented slots are averaged in the slot dimension, $\frac{1}{K} \sum_{i=1}^K \mathbf{S}_{t,k}^+$ in Eq. 8) leads to significant performance declines, underscoring the importance of these components. Using standard slot attention drastically reduces performance, highlighting the critical role of pose invariance.

2. The original SOLV results were reported at a higher resolution (336×504) with a different aspect ratio, while our experiments are conducted at 224×224 . This resolution change partly accounts for the observed performance differences.



Fig. 3. **Generation Results.** Visual comparison of video generation quality across methods on YTVIS (rows 1-2) and DAVIS17 (rows 3-4) datasets. Our method generates high-quality frames that closely match the ground truth, maintaining sharp object boundaries and preserving fine-grained textures. The results demonstrate faithful reconstruction of original scenes with superior detail preservation in the small animal’s features (row 1), natural appearance and accurate coloring of the bird (row 2), clear structural elements in the urban scene (row 3), and realistic texture and form of the white animal (row 4). Full temporal sequences are provided in the Supplementary Material.

By default, we employ 1-frame training where we randomly select one frame from each video sequence, as explained in Section 4.1. As an alternative, we evaluate 5-frame training with loss applied to all frames in the video sequences. This 5-frame training strategy yields minor performance improvements over the 1-frame approach.

As stated in Section 4.2, we omit the auxiliary attention-guidance loss in our default 1-frame training. However, when using 5-frame training, the attention guidance loss can be applied, which further improves performance but significantly increases computational costs. We therefore choose the more efficient 1-frame training as our default setting.

Comparison with Baselines. Table 2 summarizes segmentation results. We report results for three variants of SOLV: (i) encoder-only (SOLV-E), which evaluates invariant slot attention masks similar to our architecture; (ii) encoder + merging (SOLV-E+M), which applies merging to the invariant slot attention masks; and (iii) full decoder-based masks (SOLV), the default version that uses masks from the spatial broadcast decoder.

The comparison with SOLV variants reveals an important trade-off in existing approaches. While SOLV’s decoder masks achieve the highest mIoU scores, this comes at a cost: the decoder masks cannot be used for video synthesis, making the method specialized for segmentation only. When SOLV operates in configurations comparable to our approach, using encoder attention masks, its performance drops across both metrics, falling below our unified method on all metrics. Importantly, our 5-frame training nearly matches the mIoU of existing methods like SOLV, indicating its potential for further improvements, as shown in Table 1. However, we choose the 1-frame training configuration to

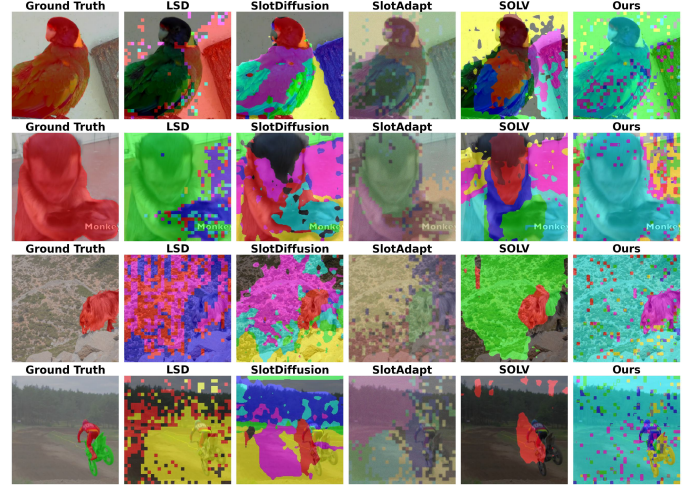


Fig. 4. **Segmentation Results.** Qualitative comparison of video object segmentation on YTVIS (rows 1-2) and DAVIS17 (rows 3-4). Our method successfully delineates objects with accurate boundaries across diverse challenging scenarios. Row 1 shows segmentation of a bird with detailed boundary preservation, row 2 demonstrates segmentation of a monkey that covers most of the frame, row 3 shows segmentation of an animal against a challenging natural background, and row 4 presents segmentation of closely touching objects (motorcycle and person) that are difficult to separate. Different colors represent distinct object instances discovered by each method. Temporal consistency across frames is demonstrated in the Supplementary Material.

balance performance with computational efficiency, representing an optimal trade-off for practical applications.

Video Generation Performance. Table 3 presents comprehensive video generation results compared to baselines. Our approach establishes new state-of-the-art results on YTVIS and DAVIS17 datasets, achieving superior performance across all complementary metrics, spanning pixel-level fidelity, structural preservation, perceptual quality, and temporal consistency.

The consistent performance across traditional pixel-based metrics (PSNR, SSIM) and more recent perceptual measures (LPIPS, FID) is particularly notable, as these metrics often exhibit trade-offs in conventional approaches. Our unified framework’s ability to simultaneously optimize for reconstruction accuracy and perceptual realism indicates a fundamental advancement in video generation quality.

The results on DAVIS17 show consistent performance across all metrics. Our method achieves lower LPIPS and FID scores compared to baselines, indicating improved perceptual quality and better distributional alignment with real video content. These improvements in perceptual metrics complement the gains observed in pixel-level measures, demonstrating the effectiveness of our approach across different evaluation criteria.

These results establish that high-quality segmentation and generation can be achieved within a unified architecture, demonstrating significant advantages over methods that specialize in single tasks. The consistent performance across both clustering-based and overlap-based segmentation metrics, combined with improvements across pixel-level fidelity and perceptual quality measures, validates our core hypothesis that temporally consistent slot representations can effectively condition temporally coherent video generation while maintaining competitive segmentation capabilities.

TABLE 3

Video generation performance on real-world datasets. We evaluate our method against state-of-the-art approaches on YTVIS and DAVIS datasets using comprehensive generation metrics. Our approach demonstrates superior performance across both pixel-level accuracy (PSNR, SSIM), perceptual quality measures (LPIPS, FID) and temporal coherence (FVD).

Method	YTVIS					DAVIS				
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow	FVD \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow	FVD \downarrow
LSD [27]	9.64	0.2793	0.777	100.68	121.41	9.58	0.0356	0.6079	84.30	75.05
SlotDiffusion [68]	9.18	0.1867	0.6484	86.38	123.8413	10.68	0.0340	0.6143	136.18	152.73
SlotAdapt [3]	10.92	0.3669	0.6556	65.30	63.72	12.18	0.0674	0.2681	41.94	29.96
Ours	11.37	0.3933	0.5908	49.51	51.77	12.38	0.0946	0.1886	28.43	16.17



Fig. 5. **Compositional Generation Results.** Demonstration of object-level editing capabilities through deletion and replacement operations on YTVIS (columns 1-3) and DAVIS17 (columns 4-5). Top rows show ground truth frames; bottom rows display edited results. Our method handles challenging scenarios including: (a) removal of closely positioned objects while preserving scene coherence (column 1, bird deletion), (b) deletion of camouflaged objects with natural background inpainting (column 3, turtle removal), and (c) semantic object replacement maintaining proper occlusion and lighting (columns 4-5). The edited videos maintain temporal consistency throughout the sequences. Full temporal sequences are provided in the Supplementary Material.

TABLE 4

Compositional generation performance. We evaluate compositional generation by mixing slots from different batch samples. Our method demonstrates superior performance across both datasets.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow
YTVIS				
LSD	8.05	0.0527	0.8734	127.49
SlotDiffusion	7.83	0.0521	0.9486	129.33
SlotAdapt	9.49	0.0575	0.7313	112.79
Ours	10.07	0.0687	0.629	83.45
DAVIS				
LSD	5.22	0.028	1.0234	198.23
SlotDiffusion	6.04	0.031	0.9912	187.13
SlotAdapt	6.93	0.032	0.9721	172.39
Ours	8.27	0.065	0.694	113.86

5.6 Qualitative Results

Segmentation Quality. Figure 4 presents visual comparisons of segmentation results across different methods on challenging video sequences. In these examples, our method successfully separates different objects with clear boundary delineation. For instance, in the bottom image of Figure 4, our approach effectively differentiates individual object instances where other methods struggle to maintain distinct segmentations.

Generation Performance. Figure 3 compares video generation results across different methods. Our method generates

higher quality frames with better detail preservation and cleaner object boundaries compared to baseline approaches. The visual comparison shows our approach maintains object identity and spatial relationships effectively while producing temporally consistent video content. For multi-frame visualizations, please refer to the Supplementary Material.

Video Editing Capabilities. Figure 5 presents examples of our framework’s compositional video editing capabilities. Thanks to a unified architecture, our method allows intuitive operations such as inserting, removing, or replacing objects, while preserving photorealistic detail. These results demonstrate the practical value of our object-centric representation for enabling flexible and interactive video editing. Additional compositional editing results are provided in the Supplementary Material.

To assess compositional generation quantitatively, we follow the setup introduced in SlotDiffusion [68] and extended in our previous work, SlotAdapt [3]. SlotDiffusion evaluates compositionality by randomly mixing slot representations from different images within a batch. We adapt this idea to the video domain by first aligning slot correspondences across frames, then randomly exchanging slots between videos in the batch, frame by frame.

As shown in Table 4, our method consistently outperforms prior approaches on this task. Taken together with the qualitative results in Figure 5, these findings confirm that our model can effectively handle compositional video

generation and editing, both in terms of visual quality and quantitative performance.

6 CONCLUSION

This work introduces the first unified framework for simultaneous video object segmentation and compositional video generation, challenging the conventional separation of these fundamental tasks. Our approach demonstrates that object-centric representations can effectively bridge perception and synthesis, achieving state-of-the-art FG-ARI performance on YTVIS and DAVIS17 datasets while establishing new benchmarks across all video generation metrics.

Our core insight is that generative modeling provides inductive structure beneficial for segmentation, while object-centric decomposition offers strong priors for temporally coherent synthesis. This synergy allows our model to outperform task-specific baselines without relying on hand-crafted temporal cues or architectural constraints. Extensive experiments confirm that integrating structured slot representations with pretrained diffusion models yields consistent improvements in temporal stability and visual fidelity.

The results demonstrate that structural commonalities between perception and generation can be exploited for mutual benefit. Strong clustering accuracy reflects improved temporal modeling, while generation metrics confirm the model's ability to synthesize content that respects both structural and semantic constraints.

Future work will explore replacing the image-based decoder with dedicated video diffusion models for end-to-end temporal dynamics modeling. Additional directions include scaling to higher resolutions and longer sequences, and incorporating text-based supervision for language-guided generation and editing.

REFERENCES

- [1] Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. Slamp: Stochastic latent appearance and motion prediction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.
- [2] Adil Kaan Akan and Fatma Güney. Stretchbev: Stretching future instance prediction spatially and temporally. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022.
- [3] Adil Kaan Akan and Yucel Yemez. Slot-guided adaptation of pre-trained diffusion models for object-centric learning and compositional generation. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2025.
- [4] Gökay Aydemir, Weidi Xie, and Fatma Güney. Self-supervised Object-centric Learning for Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [5] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: What is required and can it be learned? In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.
- [6] Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Object discovery from motion-guided tokens. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [7] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.
- [8] Ondrej Biza, Sjoerd Van Steenkiste, Mehdi SM Sajjadi, Gamaleldin Fathy Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2023.
- [9] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.
- [11] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.
- [12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.
- [13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [14] Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, and Hongkai Xiong. Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2024.
- [15] Gamaleldin Fathy Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael Curtis Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [16] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2024.
- [18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 2010.
- [19] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3-71, 1988.
- [20] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of the International Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2010.
- [21] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference.

- In *Proc. of the International Conf. on Machine Learning (ICML)*, 2019.
- [22] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.
- [23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [25] Jonathan Ho, Chitwan Saharia, William Chan, David Fleet, Mohammad Norouzi, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- [26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [27] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [29] Laurynas Karazija, Iro Laina, and Christian Rupprecht. ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation. In *In Advances of Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.
- [30] Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022.
- [31] Adam Kosior, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [32] Jannik Kossen, Karl Stelzner, Marcel Hussen, Claas Voelcker, and Kristian Kersting. Structured object-aware physics prediction for video modeling and planning. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.
- [33] Can Kucuksozen and Yucel Yemez. Hierarchical compact clustering attention (coca) for unsupervised object-centric learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [34] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [35] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.
- [36] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014.
- [38] Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2020.
- [39] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.
- [41] Anna Manasyan, Maximilian Seitzer, Filip Radovic, Georg Martius, and Andrii Zadaianchuk. Temporally consistent object-centric learning by contrasting slots. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [42] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2024.
- [43] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2013.
- [44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research (TMLR)*, 2023.
- [45] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [46] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.
- [47] Rui Qian, Shuangrui Ding, and Dahua Lin. Rethinking image-to-video adaptation: An object-centric perspective. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2024.
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2021.
- [49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image

- diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [52] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 109(5):612–634, 2021.
- [53] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.
- [54] 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. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.
- [55] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dalle learns to compose. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.
- [56] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [57] Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. Guided latent slot diffusion for object-centric learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [58] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2015.
- [59] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental Science*, 2007.
- [60] Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665, 2017.
- [61] 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.
- [62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [63] Ruben Villegas, Zalan Borsos, Aditya Ramesh, Jiahui Li, Jacob Menick, Alexander Kirillov, Oriol Vinyals, Aaron van den Oord, Nal Kalchbrenner, et al. Phenaki: Variable length video generation from open domain textual descriptions. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.
- [64] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [65] 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.
- [66] 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 *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023.
- [67] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.
- [68] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [69] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [70] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018.
- [71] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [72] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [73] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.



Adil Kaan Akan received the B.Sc. degree in computer engineering from Middle East Technical University, Ankara, Turkey, in 2020 and received the M.Sc. degree in computer science from Koc University, Istanbul, Turkey, in 2022, where he was awarded the Academic Excellence Award. He is currently a Ph.D. candidate at Koc University, supervised by Prof. Yucel Yemez. His research interests include object-centric learning, generative models, compositional image and video synthesis. His recent

work explores the integration of object-centric representations with diffusion models for controllable visual content generation.



Yucel Yemez received the B.S. degree from Middle East Technical University, Ankara, in 1989, and the M.S. and Ph.D. degrees from Boğaziçi University, Istanbul, respectively, in 1992 and 1997, all in electrical engineering. From 1997 to 2000, he was a postdoctoral researcher in the Image and Signal Processing Department of Télécom Paris (ENST). Currently, he is a Professor in the Computer Engineering Department at Koç University, Istanbul and member of the KUIS AI Center at Koç University. His research

interests include computer vision, deep learning, computer graphics, and multimedia signal processing. He has published extensively in leading journals and conferences and served as an Associate Editor for Graphical Models (Elsevier) and The Visual Computer (Springer).

APPENDIX

This supplementary material includes additional experimental details covering datasets, model configurations, and training procedures (Section A), elaborates on the Invariant Slot Attention mechanism introduced in the main paper (Section B), and provides extended qualitative visualizations (Section C).

APPENDIX A EXPERIMENTAL DETAILS

A.1 Dataset Details

To ensure consistency across inputs, we removed the black borders present in all videos from the YTVIS19 dataset. Given the self-supervised nature of our approach, we combine the standard dataset splits during training. For evaluation, we use the publicly available validation sets for all datasets except YTVIS. As the YTVIS dataset does not provide annotations for its validation split, we use the exact same subset of 300 videos from the training set that are selected by SOLV [4]. During evaluation, we upsample the predicted segmentation masks to the original frame resolution using bilinear interpolation.

A.2 Common Experimental Setup

Unless stated otherwise, all experiments use the ViT-B/14 architecture pretrained with DINOv2 [44] as the visual backbone, 7 slots with dimension 768, temporal window size $L = 5$ consecutive frames (corresponds to 2 past, 1 present, 2 future), and input resolution of 256×256 for Stable Diffusion VAE and 224×224 for the DINO encoder.

Training Schedule: We train models for 350K iterations on YTVIS19, then fine-tune for 50K iterations on DAVIS17. All models use AdamW optimizer [40] with batch size 8, gradient clipping at 0.5, and linear warmup for the first 5K iterations on $2 \times A40$ GPUs.

A.3 Model Details

Feature Extractor: We use the output of the final transformer block from DINOv2 ViT-B/14 [44], excluding the classification (CLS) token, with positional embeddings added after feature extraction.

Invariant Slot Attention: The input dimension $D_{\text{slot}} = 768$ is used throughout the ISA architecture. After positional encoding addition to DINO tokens \mathbf{F}_t , slots and projected tokens are passed to ISA. Slots are updated with a GRU cell, followed by a residual MLP with layer normalization. All projection layers (p, q, k, v, g) have dimension D_{slot} . GRU is iterated three times per frame. The scale parameter \mathbf{s}_s is multiplied by $\delta = 5$ following [4].

We initialize the components as follows: \mathbf{G}_{abs} is initialized as a coordinate grid in $[-1, 1]$, slots \mathbf{S} are initialized using Xavier initialization [20], and slot scale \mathbf{s}_s and position \mathbf{s}_p are initialized from a normal distribution [4].

Temporal Aggregator: A 3-layer, 8-head transformer encoder [62] with hidden dimension $4 \times D_{\text{slot}}$ and learnable temporal positional embeddings initialized from a normal distribution. Unavailable frame slots (indices < 0 or $>$ frame count) are masked. Unlike SOLV [4], all slots attend to one

another across frames, rather than restricting same-index slots to interact only with their corresponding slots in other frames.

Temporal Register Aggregator: A 1-layer, 8-head transformer encoder with hidden dimension $4 \times D_{\text{slot}}$. DINO features are spatially pooled before transformer input. Learnable temporal positional embeddings are initialized from a normal distribution, and unavailable frame tokens are masked.

Decoder: Adapter-injected Stable Diffusion 1.5 [50] with frozen pretrained parameters.

A.4 Baselines

All image-based baselines (LSD [27], SlotDiffusion [68], SlotAdapt [3]) are trained on flattened video frames as independent images using publicly available code implementations. To ensure fair comparison, all baselines use DINOv2 [44] as the image encoder, 7 slots with dimension 768, and follow the same training schedule of 350K iterations on YTVIS19 followed by 50K iterations fine-tuning on DAVIS17. For models requiring pretrained diffusion components, we use Stable Diffusion 1.5 [50].

SOLV [4] is trained on 5-frame sequences following the same temporal setup as our method.

APPENDIX B INVARIANT SLOT ATTENTION

This section elaborates on the mechanics of invariant slot attention (ISA), originally introduced by Biza et al. [8]. In our architecture, ISA is employed within the Object-Centric Temporal Encoding module (Section 4.1), making use of shared initialization as outlined in the main paper. Starting from the common initialization $\mathcal{Z}_t = \{(\mathbf{z}_t^j, \mathbf{s}_s^j, \mathbf{s}_p^j, \mathbf{G}_{\text{abs},t})\}_{j=1}^K$, where \mathbf{z}_t^j is the j -th slot representation, \mathbf{s}_s^j denotes the scale parameters along x and y axes, likewise, \mathbf{s}_p^j represents the position parameters for x and y axes, and $\mathbf{G}_{\text{abs},t} \in \mathbb{R}^{N \times 2}$ is the absolute coordinate grid at time t , our objective is to update the set of slots $\{\mathbf{z}_t^j\}_{j=1}^K$. To clarify, we describe here the procedure for a single time step t in the invariant slot attention mechanism:

$$\mathbf{A}_t := \text{softmax}_{j=1, \dots, K}(\mathbf{M}_t) \in \mathbb{R}^{K \times N}, \quad (11)$$

$$\mathbf{m}_t^j := \frac{1}{\sqrt{d}} p \left(k(\mathbf{F}_t) + g(\mathbf{G}_{\text{rel},t}^j) \right) q(\mathbf{z}_t^j) \in \mathbb{R}^N \quad (12)$$

Here, $q : \mathbb{R}^{D_{\text{slot}}} \rightarrow \mathbb{R}^{D_{\text{slot}}}$, $k : \mathbb{R}^d \rightarrow \mathbb{R}^{D_{\text{slot}}}$, $p : \mathbb{R}^{D_{\text{slot}}} \rightarrow \mathbb{R}^{D_{\text{slot}}}$, and $g : \mathbb{R}^2 \rightarrow \mathbb{R}^{D_{\text{slot}}}$ are learnable linear transformations applied to each patch and slot vector independently. The vector \mathbf{m}_t^j represents the j -th row of the unnormalized attention score matrix $\mathbf{M}_t \in \mathbb{R}^{K \times N}$, computing the affinity between all N spatial locations and slot j . The softmax operation is applied column-wise over the K slots, yielding the attention matrix $\mathbf{A}_t \in \mathbb{R}^{K \times N}$, where \mathbf{a}_t^j denotes the j -th row containing the normalized attention weights for slot j over all spatial locations. The relative coordinate grid associated with each slot is computed as follows:

$$\mathbf{G}_{\text{rel},t}^j := (\mathbf{G}_{\text{abs},t} - \mathbf{s}_p^j) \odot \mathbf{s}_s^j \in \mathbb{R}^{N \times 2} \quad (13)$$

where \oslash corresponds to element-wise division. The attention weights \mathbf{a}_t^j computed via (11) are used to infer both the position \mathbf{s}_p^j and scale \mathbf{s}_s^j of the slots, according to the formulation in Biza et al. [8]:

$$\mathbf{s}_s^j := \sqrt{\frac{(\mathbf{G}_{\text{abs},t}^T - \mathbf{s}_p^j \mathbf{1}_N)^2 \mathbf{a}_t^j}{\sum_{i=1}^N \mathbf{a}_t^j[i]}} \in \mathbb{R}^2, \quad (14)$$

$$\mathbf{s}_p^j := \frac{\mathbf{G}_{\text{abs},t}^T \mathbf{a}_t^j}{\sum_{i=1}^N \mathbf{a}_t^j[i]} \in \mathbb{R}^2 \quad (15)$$

where the $\sqrt{\cdot}$ and $(\cdot)^2$ operations are performed element-wise, and $\mathbf{1}_N$ is the broadcast operator that replicates the vector to match the spatial dimension (all ones row vector of dimension N). Following the attention computation, features are aggregated using a weighted combination guided by \mathbf{w}^j and projections $v: \mathbb{R}^d \rightarrow \mathbb{R}^{D_{\text{slot}}}$ and $g: \mathbb{R}^2 \rightarrow \mathbb{R}^{D_{\text{slot}}}$ applied to each patch vector independently, similarly to the original slot attention mechanism:

$$\mathbf{u}_t^j := p \left(v(\mathbf{F}_t) + g(\mathbf{G}_{\text{rel},t}^j) \right)^T \mathbf{w}_t^j \in \mathbb{R}^{D_{\text{slot}}}, \quad (16)$$

$$\mathbf{w}_t^j := \frac{\mathbf{a}_t^j}{\sum_{i=1}^N \mathbf{a}_t^j[i]} \in \mathbb{R}^N \quad (17)$$

Here, \mathbf{u}_t^j represents the aggregated features for updating slot j , and the vectors $\{\mathbf{u}_t^j\}_{j=1}^K$ form the columns of the full update matrix $\mathbf{U}_t \in \mathbb{R}^{D_{\text{slot}} \times K}$. The aggregated representations \mathbf{u}_t^j from (16) are then used to update the slot vectors $\{\mathbf{z}_t^j\}_{j=1}^K$ via a GRU module, followed by an MLP-based residual pathway as described in (18). This process is repeated over three iterative refinement steps:

$$\mathbf{z}_t^j := \text{GRU}(\mathbf{z}_t^j, \mathbf{u}_t^j) \in \mathbb{R}^{D_{\text{slot}}}, \quad (18)$$

$$\mathbf{z}_t^j := \mathbf{z}_t^j + \text{MLP}(\text{LayerNorm}(\mathbf{z}_t^j)) \quad (19)$$

Experimental Validation

We validate our register token mechanism introduced in Section 4.3. As explained in the main paper, register tokens provide spatial context to the diffusion decoder while keeping ISA slots free from spatial information, which maintains object-centric representations during generation.

Evaluation Setup. Table 5 compares our model with and without register tokens (RT) on YTVIS using the metrics from the main paper. We use the same trained model for both conditions, but replace register tokens with zero vectors during inference to isolate their effect on maintaining invariance properties.

Quantitative Results. Removing register tokens degrades performance across all metrics, confirming their role in preserving invariance. Without register tokens, reconstruction quality drops (PSNR, SSIM) because spatial information leaks into the slots, breaking their pose-invariant design. Perceptual quality also degrades (LPIPS, FID), and temporal consistency suffers significantly (FVD) as objects cannot maintain stable relationships across frames.

Temporal Analysis. The problem becomes more evident when viewing consecutive frames. Without register tokens, the same object shifts position, changes scale, and alters orientation between frames. This happens because slots now

TABLE 5

Video generation on YTVIS. Register-token (RT) ablation. The RT clearly boosts pixel accuracy (PSNR, SSIM), perceptual quality (LPIPS, FID) and temporal coherence (FVD) by carrying the pose information.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow	FVD \downarrow
w/o RT	9.90	0.28	0.69	85.0	103.0
w/ RT	11.37	0.3933	0.5908	49.51	51.77

include spatial information, which violates the invariance principle of object-centric learning. With register tokens, spatial information remains separate, so slots focus on object identity.

Visual Results. Figures 6 through 14 show results across different scenarios. Each figure displays ground truth (top), our method with register tokens (middle), and without register tokens (bottom).

Without register tokens, objects appear in wrong positions, show incorrect scaling, have inconsistent poses, and create background artifacts. These issues occur consistently across various object types and scenes.

With register tokens, results maintain correct spatial placement, consistent scaling, proper poses, and clean backgrounds. The sequences flow smoothly with natural object motion and stable spatial relationships.

These results confirm that register tokens successfully preserve ISA’s invariance properties within our diffusion framework, as detailed in Section 4.3.

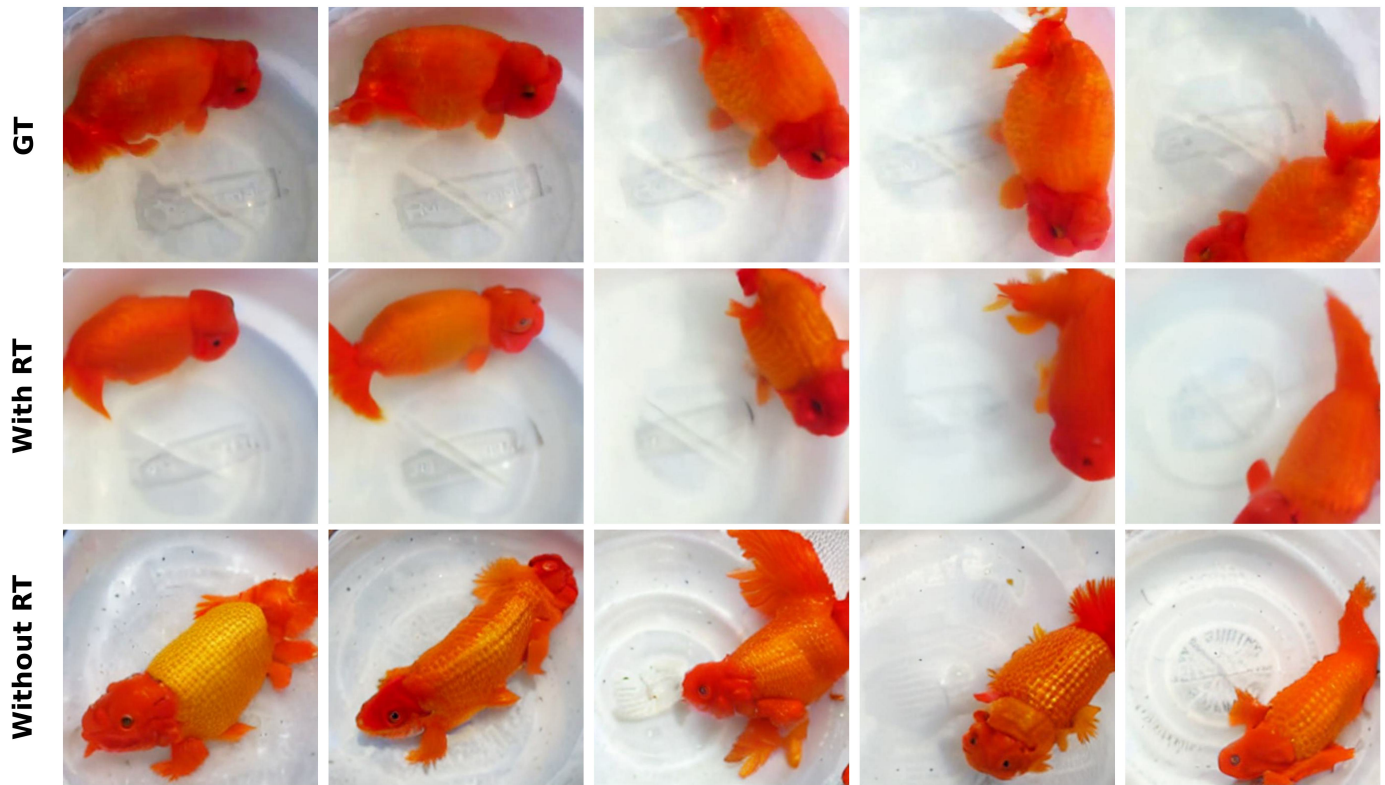


Fig. 6. Temporal video generation results with and without register tokens on YTVIS dataset. **Without register tokens** (bottom), objects appear in incorrect positions and backgrounds deviate from ground truth (top). **With register tokens** (middle), generations accurately match ground truth positioning.



Fig. 7. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).



Fig. 8. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).



Fig. 9. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).



Fig. 10. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).



Fig. 11. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).

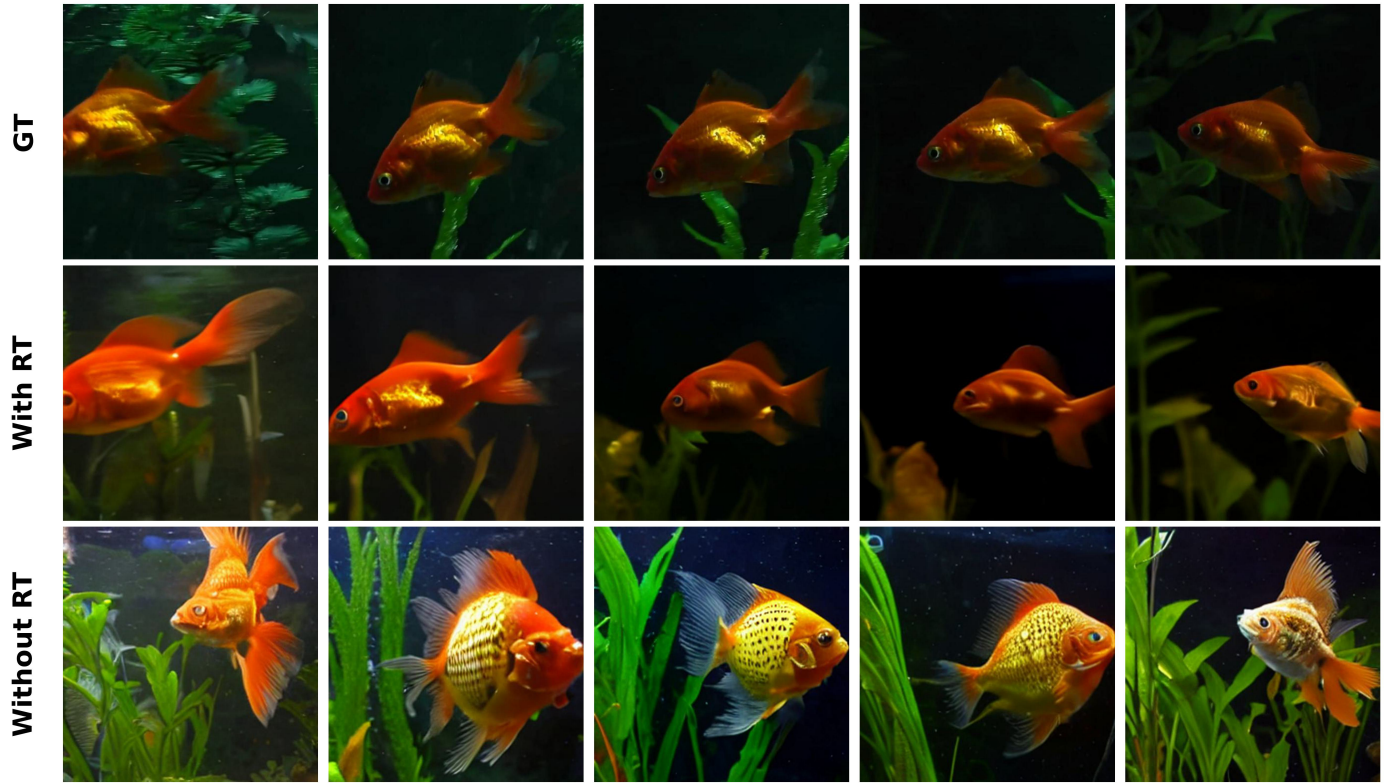


Fig. 12. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).



Fig. 13. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).



Fig. 14. Temporal video generation results with and without register tokens on YTVIS dataset. Results without register tokens (bottom) lead to spatial inconsistencies across frames. Results with register tokens (middle) maintain consistent object positioning relative to ground truth (top).

APPENDIX C

COMPARISON WITH BASELINES

This section presents comprehensive video generation and segmentation results, comparing them against baseline methods across temporal sequences. While single-frame results are provided in the main paper, this supplementary material emphasizes temporal consistency and visual fidelity across consecutive frames to demonstrate the effectiveness of our unified slot-based framework.

C.1 Generation Results

Figures 15, 16, 17, and 18 demonstrate our method’s video generation capabilities across multiple consecutive frames on YTVIS and DAVIS17 datasets. Each figure shows five temporal frames from a single video sequence, with each row representing a different time step. The leftmost column shows ground truth frames, followed by results from baseline methods (LSD, SlotDiffusion, SlotAdapt), and our method in the rightmost column.

Our approach demonstrates improved temporal coherence, object identity preservation, and visual fidelity throughout the sequences. Key aspects to observe include: (1) structural stability of objects across time, (2) consistency of fine-grained details such as textures and colors, (3) natural motion dynamics, and (4) preservation of spatial relationships between objects and backgrounds. Baseline methods typically exhibit temporal artifacts, inconsistent object representations, and degradation in visual quality over time, while our unified framework successfully handles these challenging scenarios through effective slot-based temporal binding. The results show our method’s capability to generate high-quality, temporally consistent video content that maintains object coherence across complex motion patterns.

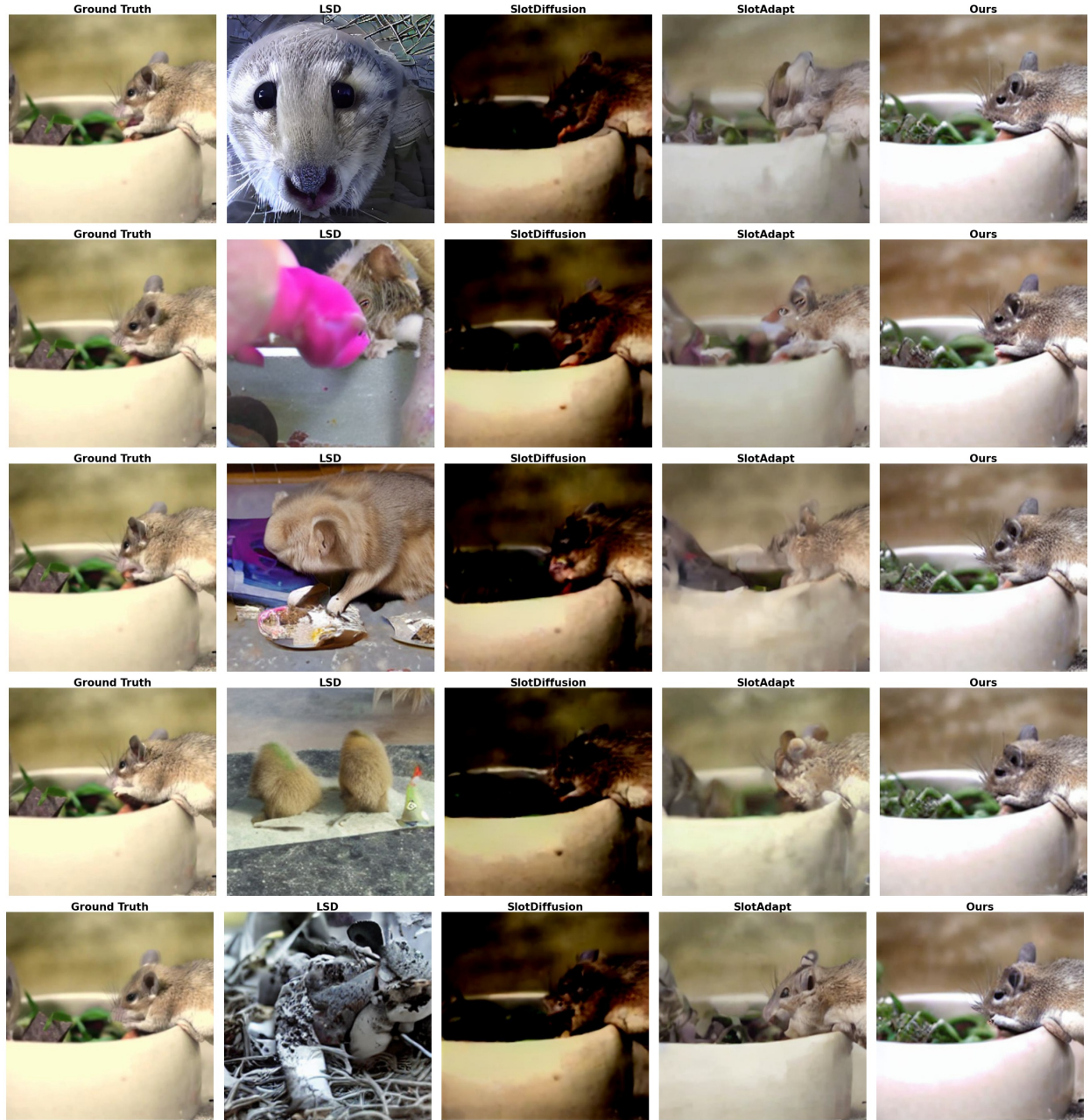


Fig. 15. **Temporal Video Generation on YTVIS Dataset.** Multi-frame generation results showing object identity preservation and spatial coherence across five consecutive frames.



Fig. 16. **Temporal Video Generation on YTVIS Dataset.** Video generation on challenging sequences with multiple objects and complex backgrounds.



Fig. 17. **Temporal Video Generation on DAVIS17 Dataset.** Results demonstrating object motion tracking and spatial layout consistency over time.



Fig. 18. **Temporal Video Generation on DAVIS17 Dataset.** Generation results highlighting fine-detail preservation throughout temporal sequences.

C.2 Segmentation Results

Figures 19, 20, 21, and 22 demonstrate our method’s unsupervised video object segmentation performance across temporal sequences on YTVIS and DAVIS17 datasets. Each figure displays five consecutive frames with segmentation masks overlaid, where different colors represent distinct object instances discovered by each method. The evaluation focuses on temporal binding consistency—the ability to maintain stable object identity and accurate boundaries across frames.

Our slot-based representations successfully handle various challenging scenarios including: (1) rapid object motion and deformation, (2) objects with similar appearances or spatial proximity, (3) scale changes and partial occlusions, and (4) complex multi-object interactions. The results demonstrate our framework’s improved capability in preserving object identity and spatial coherence compared to baseline approaches, which often exhibit segmentation instability, identity confusion, and boundary degradation across temporal sequences. The consistent performance across diverse video content shows the effectiveness of our temporal object-centric learning approach for video understanding applications.

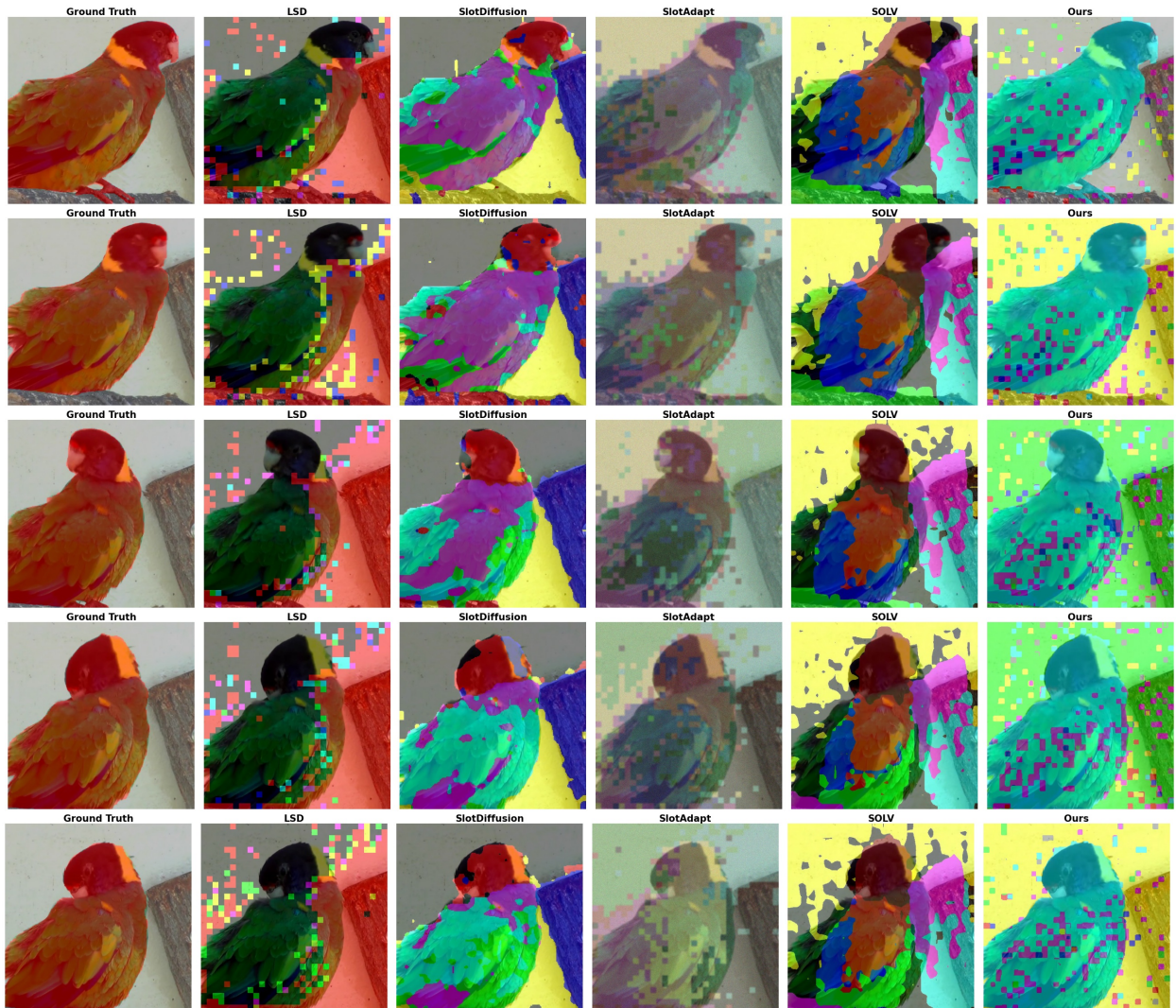


Fig. 19. **Segmentation Results on YTVIS.** Consistent object boundary detection and identity preservation across dynamic motion and pose changes.

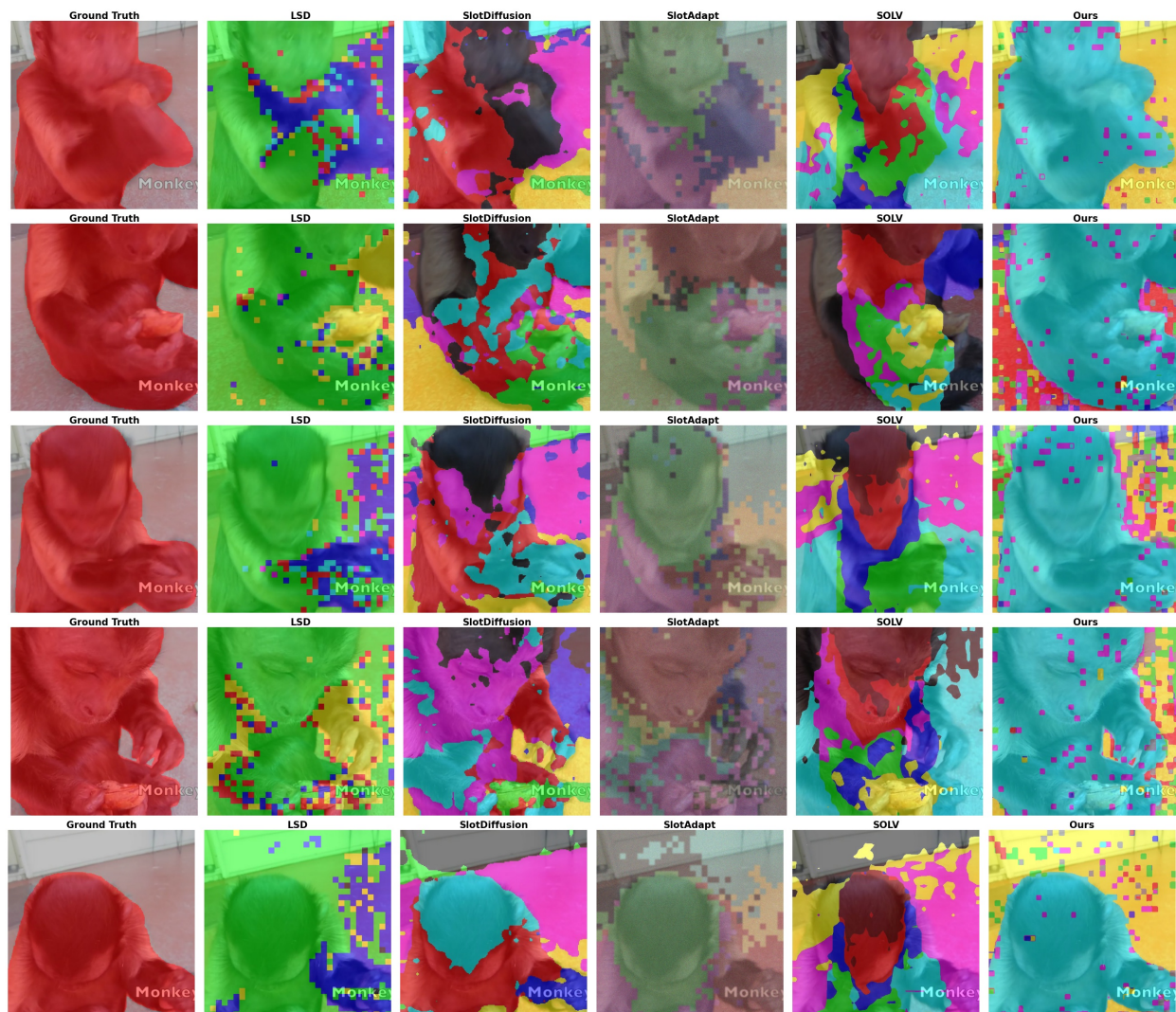


Fig. 20. **Segmentation Results on YTVIS.** Handling of significant shape and appearance variations with temporal tracking.

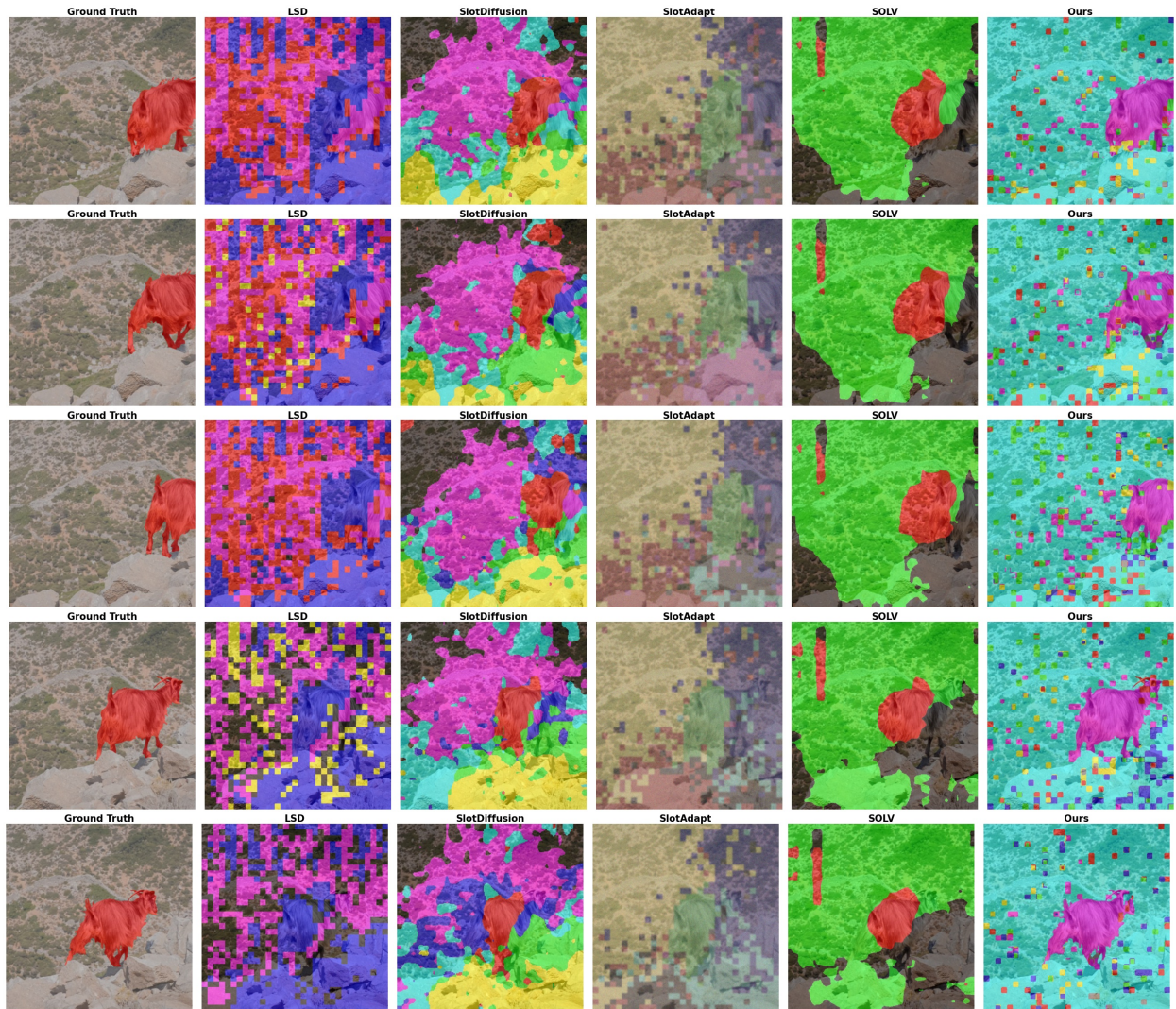


Fig. 21. **Segmentation Results on DAVIS17.** Tracking and segmentation of small, fast-moving objects through rapid motion and scale changes.

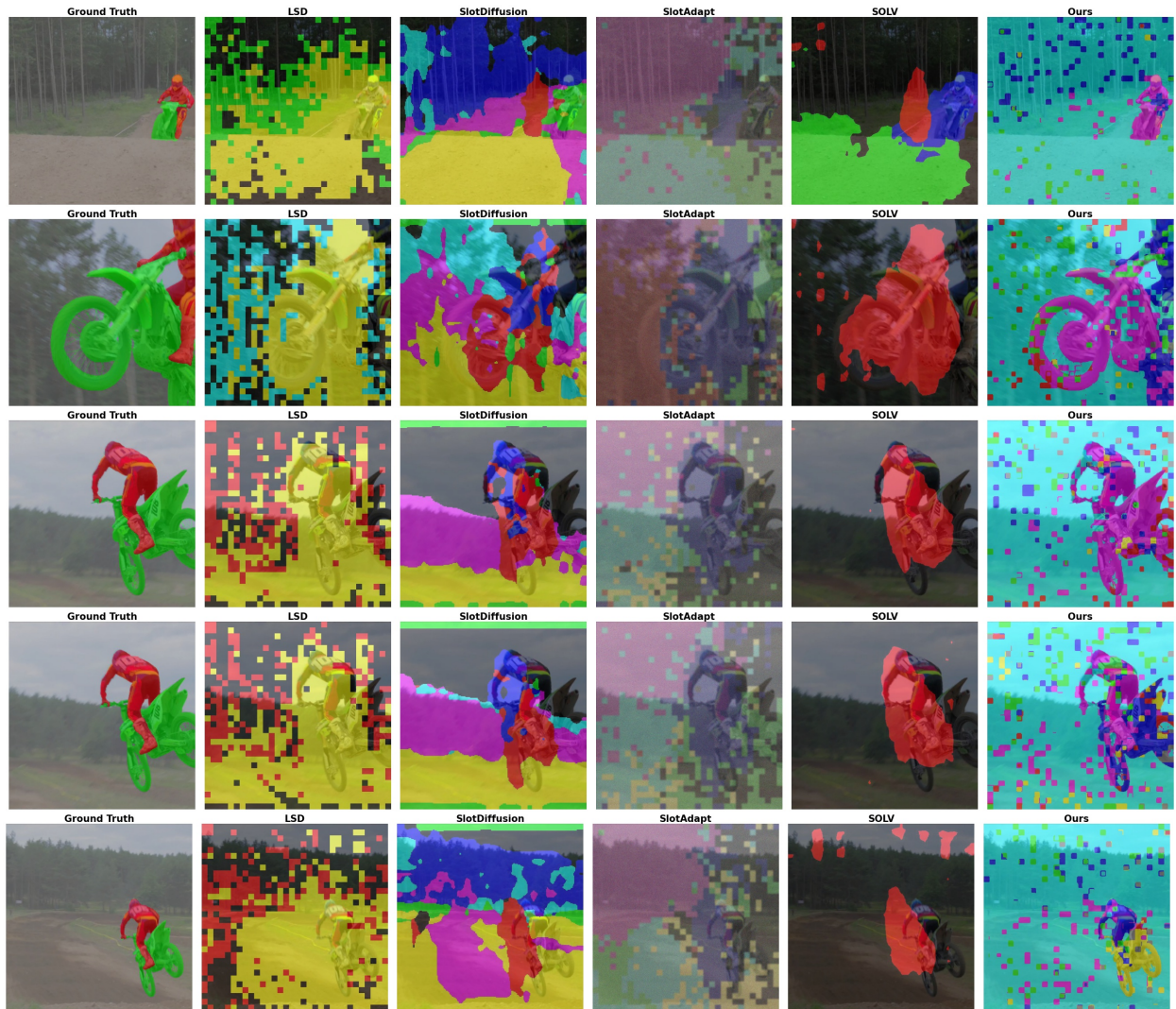


Fig. 22. **Segmentation Results on DAVIS17.** Distinguishing and tracking multiple closely positioned objects with stable segmentation masks.

C.3 Compositional Generations

Figures 23, 24 and 25 demonstrate our framework’s compositional editing capabilities through object deletion and replacement across temporal sequences. Our slot-based representation enables targeted removal or replacement of specific objects while maintaining scene coherence and temporal consistency. The top row (GT) shows original ground truth frames, while the bottom row (Gen) displays generated results after object removal.

The model addresses several technical challenges: (1) background inpainting where objects were removed, (2) preservation of natural motion dynamics in remaining scene elements, (3) maintenance of lighting and shadow consistency, and (4) integration of filled regions with surrounding context. These results show our approach’s effectiveness in enabling object-level control for video editing applications, demonstrating the practical utility of our unified slot-based framework for compositional video manipulation tasks.

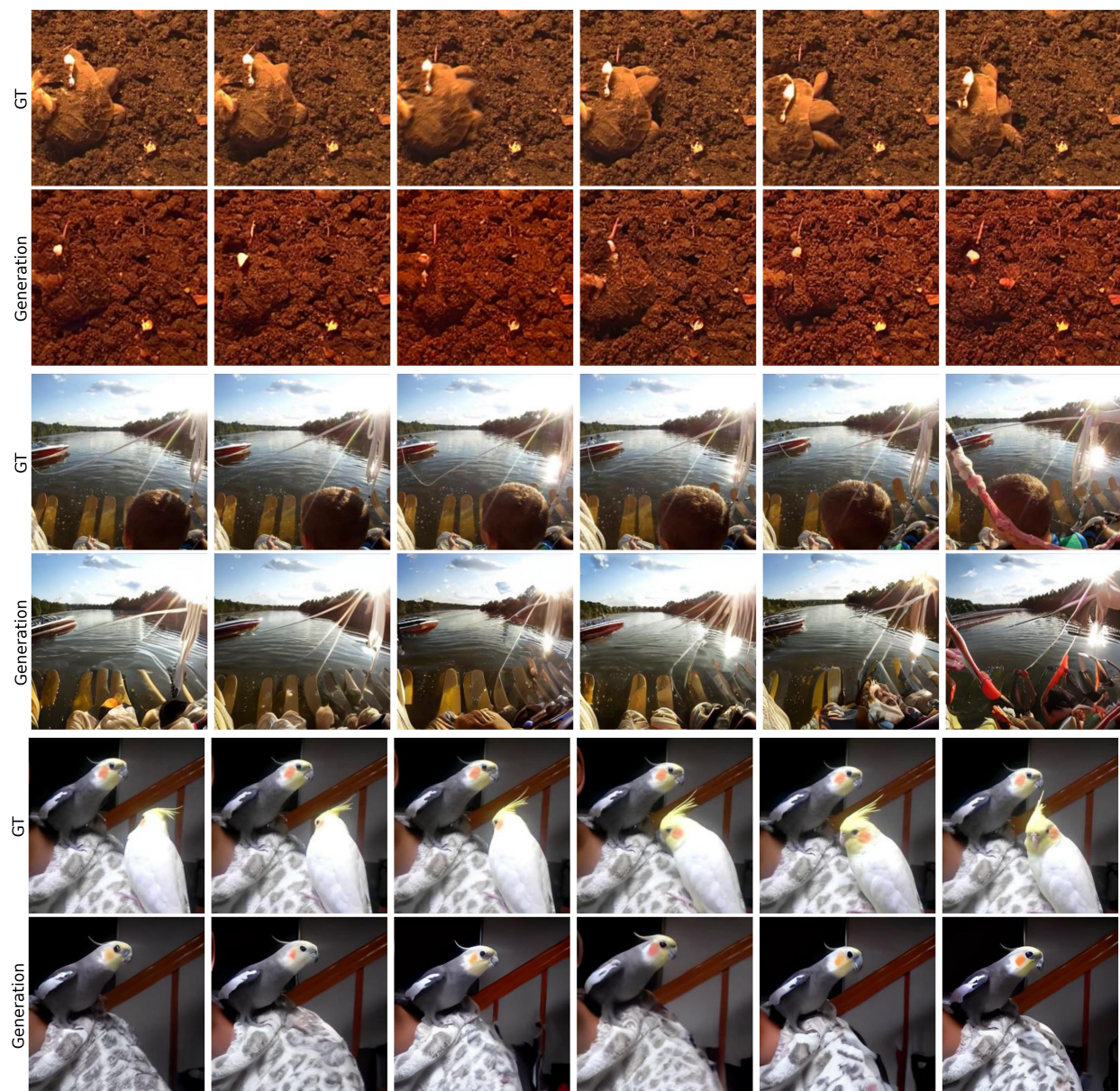


Fig. 23. **Compositional Editing Examples.** Targeted object removal while maintaining scene coherence and temporal consistency.

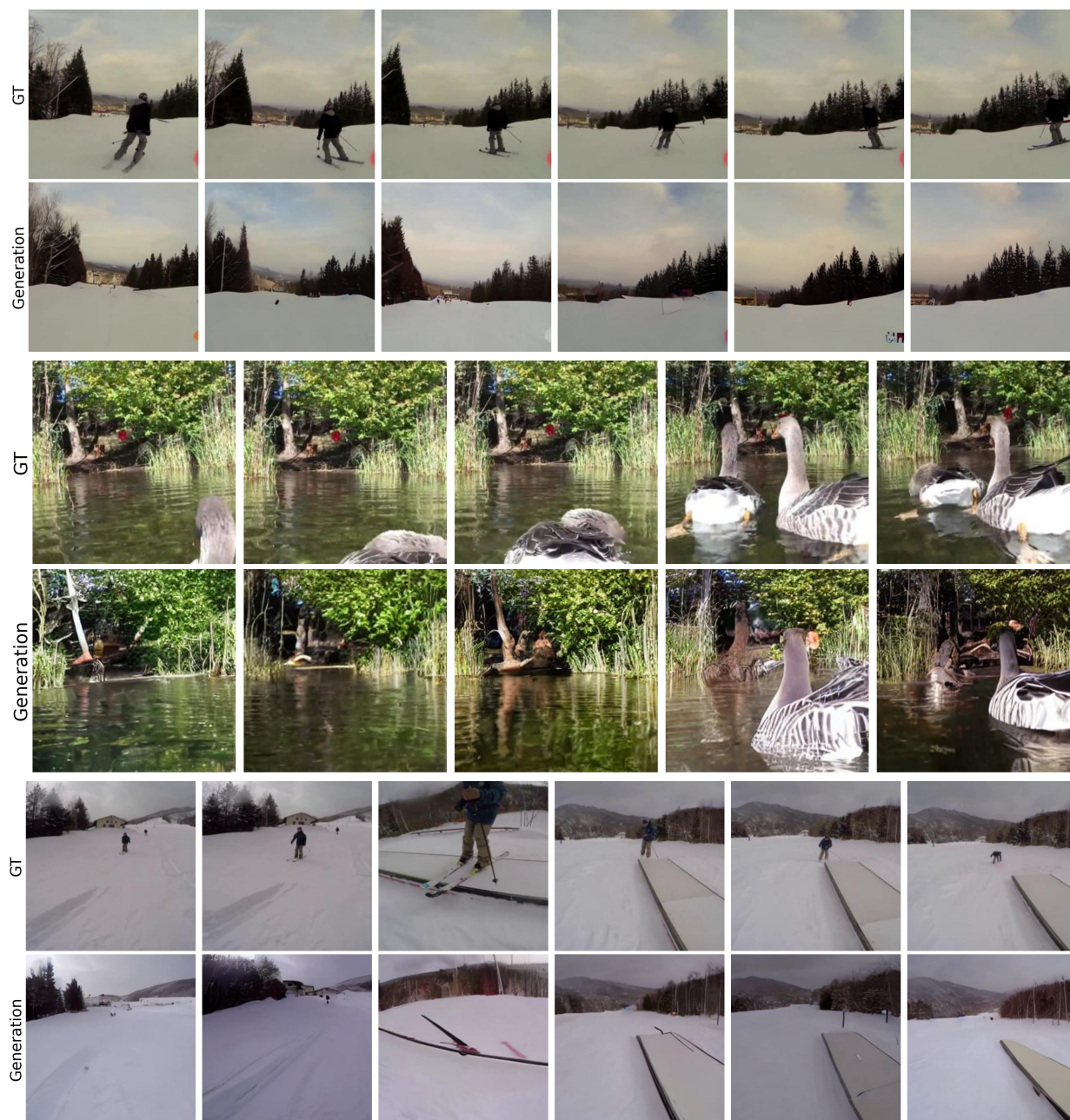


Fig. 24. **Compositional Editing Examples.** Object deletion across diverse video sequences with realistic scene completion.

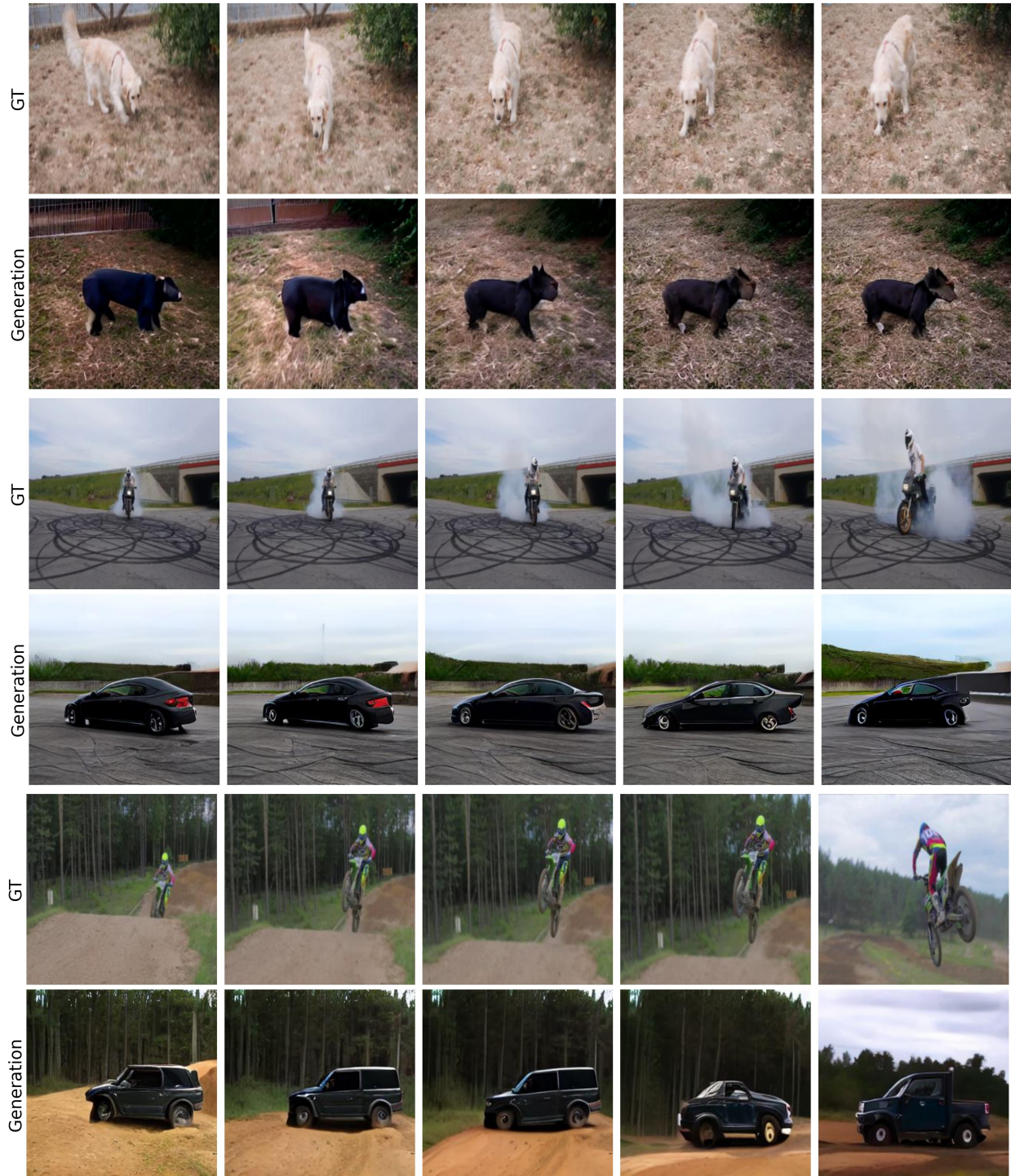


Fig. 25. **Compositional Editing Examples.** Targeted object replacement is performed while preserving scene coherence and temporal consistency. *Top:* the white dog is replaced with a black dog. *Middle and bottom:* the motorcycle is replaced with a black car.