International Journal of Applied Sciences, Research, and Innovation

OPEN ACCESS

FULL PAPER | IJASRI | OPEN

A Layered Spectral Memory Framework for Reference Object Recasting in Source Preserving Video Transformation Under Occlusion

Mohamed Trabelsi[1], **Walid Gharbi**[2], **Nizar Jlassi**[3]

[1]Department of Computer Science, University of Jendouba, Avenue de l'Union du Maghreb Arabe, Jendouba 8189, Tunisia
[2]Department of Information Technology, University of Kairouan, Rue Assad Ibn Al Fourat, Kairouan 3100, Tunisia
[3]Department of Computer Science, University of Gafsa, Cité Zrig, Gafsa 2112, Tunisia

Abstract

Video object editing often requires an inserted visual identity to behave as though it belonged to a source recording rather than a separately generated image sequence. The difficulty is amplified when a user provides only one reference image, since the system must infer unseen views, occluded parts, and contact behavior from limited evidence. This paper introduces Layered Spectral Memory Recasting, a source preserving video transformation framework that treats reference based editing as a layered transport problem rather than a framewise synthesis task. The method separates a video into a carrier layer, an interaction layer, and a reference layer, then uses a spectral memory bank to retain reference-specific color, material, and part-frequency structure across time. A recurrent transport ledger propagates editable support through source motion while an occlusion ledger suppresses reference hallucination during disappearances and reappearances. The final generator operates in a discrete latent token space with spectral residual constraints, allowing object identity to remain coherent without forcing the replacement to inherit every contour of the source object. Experiments on a constructed benchmark of 126 short videos and 252 single-image replacement requests show that the framework improves reference retention by 10.8%, reduces boundary contamination by 15.6%, and lowers reappearance identity drift by 19.2% against strong token, atlas, and frame-conditioned baselines. The results indicate that reference-object recasting benefits from separating temporal transport, boundary negotiation, and reference memory rather than merging them into a single conditioning pathway.

1. Introduction

A single image can specify an object with more visual precision than a phrase can usually carry. A product photograph fixes a shoe's panel arrangement, a ceramic mug's glaze, a toy's painted face, or a vehicle's side proportions. Video editing systems, however, do not receive such a photograph as a complete object model. They receive a view, a texture sample, and a set of ambiguous hints. When the source video moves, turns, blurs, or hides the object behind another surface, the replacement must continue as a plausible temporal entity. This is the central problem addressed here: how to recast an object in a source video using one reference image while preserving the scene's temporal and spatial behavior.

The usual formulation of video object editing places a large burden on a generative model. The model is expected to infer where the object should exist, how its appearance should change across frames, how it should pass behind occluders, and how its texture should resist frame-to-frame drift. A highly capable generator can sometimes solve these tasks implicitly, but practical failures often show that the tasks are not the same. Region placement, contact handling, material consistency, and identity retention fail in different ways. A method that blurs these functions into one broad conditioning operation is hard to diagnose and hard to repair. This paper therefore adopts a decomposed view of reference-object editing.

The proposed framework, Layered Spectral Memory Recasting, treats the source video as a set of interacting layers. The carrier layer contains background structure and camera behavior. The interaction layer contains hands, shadows, contact surfaces, partial occluders, and nearby objects that should constrain the replacement. The reference layer contains the generated object. These layers are not assumed to be perfectly segmented. They are soft computational roles used to decide which signal should dominate at each stage of editing. A spectral memory bank stores reference-specific appearance statistics in a way that is less sensitive to pose than raw pixels. A recurrent transport ledger estimates how editable support moves through time. An occlusion ledger records when the object should be hidden, partially visible, or reintroduced after disappearance.

This design is motivated by a simple observation: video editing errors often arise after a temporary loss of evidence. A replacement may look reasonable in visible frames, vanish behind an arm, and then return with a different stripe pattern or object scale. A framewise system has no memory of what was hidden. A purely flow-based system can preserve motion but may drag stale texture through occlusion. A full generative video model may create a smooth result but weaken

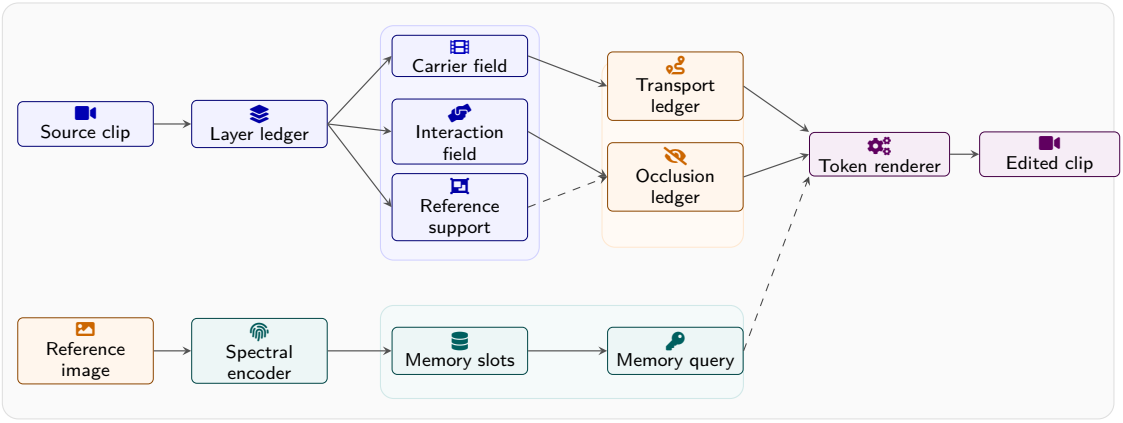


Figure 1. Layered recasting pipeline separates carrier timing, interaction constraints, reference support, spectral memory, occlusion state, and token rendering into inspectable modules.

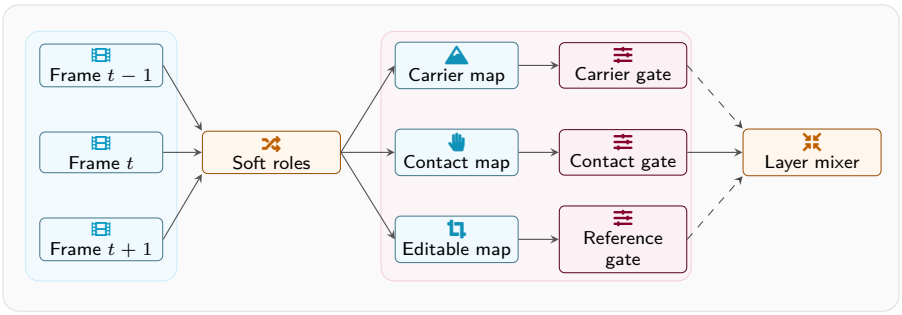


Figure 2. Layer ledger converts adjacent frames into soft carrier, contact, and editable maps before gated mixing, reducing boundary conflict during local object replacement.

the reference identity. The present work asks whether a separate memory mechanism can help the generator treat the reference object as persistent, even when the rendered frame contains only partial evidence.

The framework differs from conventional video-to-video translation in its reliance on a single external object image. Paired translation methods learn mappings between domains and can synthesize temporally structured outputs when training data matches the target task [19, 32]. Neural rendering methods can preserve detailed appearance when scene geometry or learned texture fields are available [29, 21]. Dynamic radiance approaches can model nonrigid scenes, but they normally require multiview or dense temporal evidence that a single target photograph does not provide [22, 25]. Reference-object recasting must work in a weaker information regime: one source clip, one target image, and no complete object geometry.

Harsha et al. [15] showed that a target image can serve as a useful control signal for video editing when text alone cannot specify object shape and instance appearance; that result is significant for the present study because it locates the reference image inside the editing problem rather than treating it as optional decoration. The present paper takes a different route by emphasizing layered transport and spectral memory. Instead of centering the design on one denoising pathway, it asks how a persistent reference record can be maintained while the video passes through occlusion, blur, and boundary contact.

The framework also draws on a longer tradition of visual completion and layered manipulation. Early image inpainting methods treated missing regions through geometric continuation and exemplar search [3, 8]. Learned inpainting systems later introduced semantic priors for filling absent content [23, 36]. Video editing adds temporal constraints to this completion problem, and object replacement adds a reference identity that must survive across time. The present work can be viewed as a reference-bound completion method where the absent region is not simply unknown background but a new object layer constrained by source motion.

The paper makes three technical contributions. First, it defines a layer ledger that distinguishes carrier, interaction, and reference signals during editing. Second, it introduces a spectral memory bank that stores reference appearance as multiscale frequency descriptors and discrete visual tokens. Third, it evaluates an occlusion-aware recurrent transport mechanism that controls when the replacement should be propagated, paused, or regenerated. These elements are deliberately modular, so that each failure mode can be inspected separately rather than hidden inside the final visual output.

The experimental setting is narrow but demanding. The source clips are short videos with object motion, partial occlusion, and background contact. The replacement object comes from a single photograph. The intended edit is not a full style transfer or an unconstrained video generation; it is a local recasting of one object while retaining source timing and scene structure. The results are reported with ref-

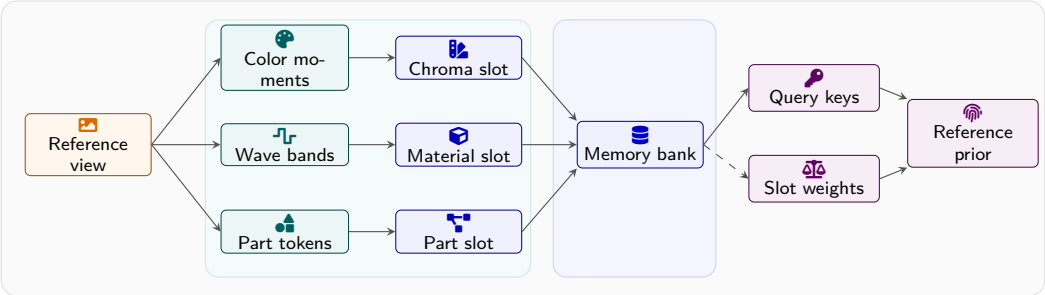


Figure 3. Spectral memory stores reference appearance as color statistics, multiscale frequency bands, and part tokens, then retrieves compact priors during visible rendering and reappearance.

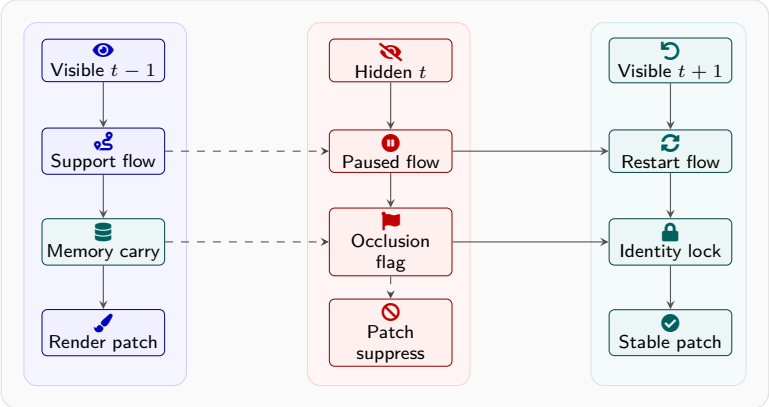


Figure 4. Occlusion-aware recurrent transport pauses editable support during disappearance, preserves reference memory, and relocks identity when the object becomes visible again.

erence retention metrics, boundary contamination measures, reappearance consistency, and human preference judgments. The findings suggest that temporal object identity is improved when the generator has access to a durable reference memory and a separate occlusion state, rather than relying only on per-frame image conditioning.

2. Related Work and Positioning

Generative modeling for visual editing has moved through several distinct regimes. Generative adversarial networks established a powerful framework for learning image distributions and enabled early high-quality image synthesis [13]. Conditional adversarial systems made image translation more controllable, especially when paired data or structured labels were available [19]. Video-to-video synthesis extended this logic to temporal data, using learned temporal constraints to improve output continuity [32]. These methods remain relevant because they frame editing as conditional transformation rather than unrestricted creation. Their main limitation for the present setting is that a one-shot reference object does not define a full translation domain.

Transformer and token-based generative models provide another foundation. Vision transformers established that global self-attention can serve as a general visual representation mechanism [9]. Video transformers such as TimeSformer and ViViT adapted attention over space and time, allowing long-range relationships to enter action and video understanding models [4, 1]. Multiscale vision transformers introduced hierarchical attention structures suitable for efficient video representation [10]. Masked token generation later showed that visual content can be generated by iteratively filling discrete latent tokens rather than by direct pixel regression [7]. The present work uses a discrete token generator because token replacement allows memory retrieval and reference constraints to be expressed at the level of reusable visual units.

Diffusion and autoregressive video generators have advanced text-conditioned video synthesis, but they often pursue broad generative coverage rather than precise local recasting. CogVideo explored large-scale text-to-video generation with transformer structures [18]. Phenaki addressed variable-length video generation from open-domain prompts [30]. Imagen Video developed cascaded video diffusion models for high-resolution text-to-video synthesis [17]. MagicVideo introduced efficient latent video generation procedures [37]. VideoComposer incorporated multiple spatial and temporal conditions into video generation [33]. These systems are important because they show that temporally aware generative priors can synthesize motion, but local replacement from a single reference image requires tighter source preservation and a different treatment of identity memory.

Video completion and conditional prediction methods are also relevant. Flexible diffusion modeling of long videos examined generation under partially observed temporal contexts [16]. MCVD studied masked conditional video diffusion for prediction and interpolation, demonstrating the value of condi-

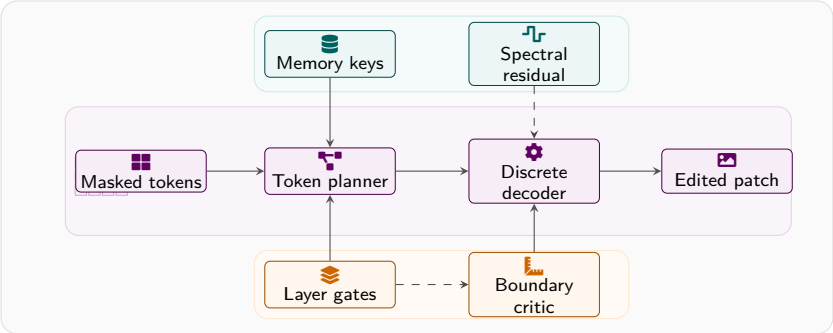


Figure 5. Discrete latent rendering fills masked object tokens under memory-key retrieval, layer gating, spectral residual constraints, and boundary criticism.

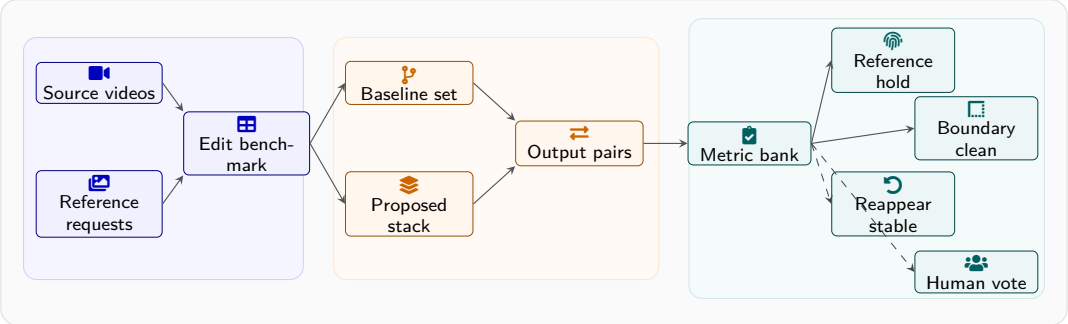


Figure 6. Evaluation pipeline compares edited outputs with reference retention, boundary contamination, reappearance drift, and human preference measures; the reported gains are 10.8%, 15.6%, and 19.2% on the three automatic criteria.

tioning on visible context [31]. Neural atlas methods decompose video into canonical layers, allowing edits to be applied through a persistent representation rather than independently across frames [20]. Text2LIVE used layered neural atlases for text-driven video layer editing [2]. These works provide useful precedents for layer-based thinking, but reference-object recasting differs because the inserted object is not merely a transformed version of an existing layer.

Tracking and segmentation methods shape the practical side of the problem. The DAVIS benchmark made semi-supervised video object segmentation a central evaluation setting [24]. YouTube-VOS expanded video object segmentation to a larger and more varied dataset [34]. OSVOS demonstrated that a single annotated frame can guide object segmentation through video [6]. AOT introduced an association-based architecture for tracking and segmenting multiple objects through time [35]. For the present task, segmentation is not the final answer because the replacement may occupy regions absent from the source object. Still, segmentation helps identify interaction zones and occlusion events.

Optical flow and geometry estimation offer another set of tools. The KITTI benchmark foregrounded realistic driving scenes for flow and stereo estimation [11]. The Sintel benchmark offered challenging nonrigid motion and cinematic effects [5]. EpicFlow used edge-aware interpolation to refine sparse matches [27]. PWC-Net showed that pyramids, warping, and cost volumes can be combined effectively for learned optical flow [28]. Monocular depth methods such as unsupervised depth and pose learning, Monodepth2, and MiDaS provide approximate scene structure useful for contact and occlusion reasoning [38, 12, 26]. The proposed framework uses flow and depth only as advisory signals because reference-object geometry cannot be recovered fully from one image.

The present paper sits at the intersection of these strands but does not claim to replace them. It borrows the layered view from atlas methods, the discrete representation from token generators, the temporal awareness of video transformers, and the boundary discipline of segmentation and flow systems. Its main distinction is the explicit spectral memory assigned to the reference object. Rather than repeatedly asking a generator to rediscover the reference appearance, the framework carries a compact record of appearance, part frequency, and token co-occurrence throughout the video. The generator then consults this record during visible rendering and reappearance events.

3. Layered Spectral Memory Recasting

The editing problem is written as a transformation from a source clip $V = \{x_1, \dots, x_T\}$ and a reference image r to an edited clip $Y = \{y_1, \dots, y_T\}$. The source clip contains an object region that is to be recast. The reference image contains the desired replacement object. The system is not given a 3D model, a target-video example, or additional photographs of the replacement. The task is therefore underdetermined. The framework handles this by separating the problem into layer assignment, temporal transport, reference memory retrieval, and token-space rendering.

The layer ledger contains three fields for each frame. The car-

rier field C_t describes pixels whose appearance should remain close to the source. The interaction field I_t describes regions that may not be replaced but can constrain the replacement, such as fingers, shadows, floor contact, vehicle tires near the road, or a strap crossing an object. The reference field R_t describes where the new object may be rendered. These fields are soft maps rather than crisp masks. They can overlap when the system is uncertain. For example, a hand holding a bottle may occupy the interaction field while the bottle itself occupies the reference field. At the boundary, the same pixels may have nonzero carrier and reference weights until rendering resolves the final transition.

The fields are initialized from segmentation, motion, and depth cues. A segmentation model produces candidate object regions, flow links the candidates across time, and a depth estimator identifies likely foreground occluders. The initialization is intentionally conservative around interaction zones. A replacement should not erase a hand merely because the reference object is visually strong. The layer ledger is then refined recurrently. At each frame, the model predicts whether a location should preserve source appearance, accept reference rendering, or behave as a mixed contact boundary.

The reference image is encoded into a spectral memory bank. The memory bank has two parts. The first part stores multiscale frequency statistics from the object crop. A Laplacian pyramid and a blockwise Fourier transform produce descriptors for color bands, edge orientation, repeated local patterns, and low-frequency material tone. The second part stores discrete visual tokens from a learned vector-quantized encoder. These tokens capture part-like appearance patterns without requiring a semantic part detector. The memory is not a copy-paste buffer. It stores how the reference looks at multiple scales so that the generator can recreate consistent evidence under altered pose and partial visibility.

Let $\mathcal{P}_k(r)$ denote the k-th pyramid band of the reference crop and $\mathcal{F}$ denote a local Fourier transform. The spectral record is

$$S_{k,b} = \left[|\mathcal{F}_b(\mathcal{P}_k(r))|, \quad \angle \mathcal{F}_b(\mathcal{P}_k(r)) \right],$$
$$\mathcal{S} = \{S_{k,b}\}_{k=1,b=1}^{K,B}.$$

The absolute and phase components are stored separately because material color and repeated patterns behave differently under deformation. Low-frequency magnitude is relatively stable across view changes, while phase-like descriptors can be unreliable under rotation and occlusion. The generator learns gates that decide which components to trust at each frame.

The discrete token record is produced by a vector-quantized encoder q_ψ. For the reference crop, the encoder returns token indices u^r and continuous features h^r. A token co-occurrence graph is constructed from neighboring tokens. This graph helps prevent local identity drift. If a generated wheel-like token appears next to a body-color token in one frame, the memory graph encourages similar adjacency in nearby frames unless the occlusion ledger says that the region is hidden. This is particularly useful for objects with repeated patterns or logos, where single-token similarity is insufficient.

The transport ledger estimates how the reference layer moves. It does not attempt to recover exact physical motion of the target object. Instead, it predicts a carrier-aligned support map that says where reference tokens may be placed. Source flow provides an initial motion cue, but the ledger can deviate from source flow when the replacement is larger or has different proportions. This is accomplished through a learned offset field d_t defined over the reference field. The field is regularized by local smoothness and by interaction boundaries.

The recurrent state update is

$$\begin{aligned} m_t &= \Phi_m(m_{t-1}, \quad f_{t-1\to t}, \quad C_t, \quad I_t, \quad R_t), \\ d_t &= \Phi_d(m_t, \quad z_{t-1}, \quad z_t^c), \\ \hat{R}_t &= \mathcal{W}(R_{t-1}, \quad f_{t-1\to t} + d_t). \end{aligned}$$

Here m_t is the memory state, $f_{t-1\to t}$ is source flow, z_{t-1} is the prior rendered token field, z_t^c is a carrier feature for the current frame, and $\mathcal{W}$ is a differentiable warp. The offset field allows the support to follow source motion while not being locked to the source object's exact outline.

The occlusion ledger is a binary-relaxed state variable $o_t(p)$ that estimates whether a reference-layer location is visible, partially hidden, or absent. Occlusion is inferred from depth discontinuities, flow inconsistency, segmentation changes, and interaction-field overlap. If the object passes behind a pole, hand, or foreground surface, the occlusion ledger suppresses token regeneration in the hidden region. The memory state remains active, but rendering is delayed until visibility returns. This prevents a common error in which a generator fills an occluder with reference texture because it is trying too hard to maintain identity.

The rendering generator operates in a discrete latent grid. It receives carrier tokens from the source frame, reference tokens from the memory bank, and layer weights from the ledger. It predicts edited tokens, which are decoded into RGB frames. Discrete tokens are convenient because the generator can choose among memory-consistent appearance units rather than synthesizing every local patch continuously. The rendering process is inspired by masked token modeling, but it is not a generic image completion task. The mask to be filled is the reference field, and the fill operation is constrained by spectral memory, interaction boundaries, and temporal state.

The token prediction objective is

$$\mathcal{L}_{tok} = -\sum_{t,p} R_t(p) \log P_\theta(u_{t,p}^y \mid u_t^c, \mathcal{M}, m_t, o_t),$$
$$\mathcal{M} = (\mathcal{S}, u^r, h^r, G^r),$$

where $u_{t,p}^y$ is the edited token at position p, u_t^c is the carrier-token grid, $\mathcal{M}$ is the reference memory, and G^r is the token co-occurrence graph. During training, edited-token targets are obtained from synthetically constructed recasting examples and from pseudo-replacements created with controlled compositing. During inference, the model samples tokens iteratively and uses confidence scores to revisit uncertain regions.

The spectral residual loss compares reference-consistent frequency content in the generated reference layer. It is not applied directly to the whole object at full strength because the reference may appear under a different viewpoint or lighting. Instead, the loss compares selected statistics over visible high-confidence regions. It penalizes loss of stable color bands and repeated local patterns while allowing phase shifts

at boundaries.

$$\mathcal{L}_{spec} = \sum_{t,k,b} w_{t,k,b} \left\| A_{k,b}(y_t \odot R_t) - A_{k,b}(r) \right\|_1 + \sum_{t,k,b} \tilde{w}_{t,k,b} \left\| P_{k,b}(y_t \odot R_t) - P_{k,b}(r) \right\|_1 .$$

Here A and P denote magnitude and phase descriptors. The weights w and $\tilde{w}$ come from visibility and pose uncertainty. Phase penalties are reduced when the system estimates strong viewpoint change.

The boundary term protects the carrier and interaction fields. It discourages reference color from entering nearby hands, roads, tabletops, or background objects. The loss uses a narrow band around the reference field and compares the edited frame with the source frame after excluding pixels that are allowed to change. This is a practical compromise. Perfect physical relighting would alter some boundary pixels, especially shadows, but uncontrolled boundary spread is more damaging for local recasting.

The complete training objective is

$$\mathcal{L} = \lambda_{tok}\mathcal{L}_{tok} + \lambda_{spec}\mathcal{L}_{spec} + \lambda_{car}\mathcal{L}_{car} + \lambda_{occ}\mathcal{L}_{occ} + \lambda_{edge}\mathcal{L}_{edge} + \lambda_{temp}\mathcal{L}_{temp}.$$

The carrier term preserves source appearance outside the reference layer. The occlusion term trains the visibility ledger. The edge term reduces haloing. The temporal term measures token and feature consistency after transport. The terms are not equally active during all training phases. Early training emphasizes token reconstruction and carrier preservation. Later training increases spectral and occlusion weights.

The framework's operational logic can be summarized as follows:

- The source video supplies motion, carrier appearance, contact surfaces, and occlusion cues, but it does not dictate the final object silhouette.
- The reference image supplies material and identity evidence through spectral descriptors and token co-occurrence rather than through repeated raw-image injection.
- The layer ledger decides which locations are edited, protected, or treated as contact zones, and this decision can change through time.
- The occlusion ledger prevents the generator from inventing visible reference texture where the object should be hidden.

This organization creates an editing process that is less dependent on a single all-purpose mask. The reference layer can grow beyond the source object when needed, but the interaction layer can still protect occluding hands or foreground barriers. The memory bank can preserve a distinctive object pattern, but the spectral gates can reduce overcommitment when the view changes. The token generator can synthesize unseen details, but the co-occurrence graph discourages sudden local rearrangements. These are small design choices, yet they collectively address several recurring problems in one-image video replacement.

4. Training Data Construction and Inference Procedure

Training a reference-object recasting model requires examples where a source object, a reference object, and an edited output are known. Such data is rarely available at scale. The training strategy therefore combines three sources of supervision. The first source is synthetic layer composition, where segmented objects are inserted into unrelated videos with controlled motion and occlusion. The second source is self-recasting, where an object from one frame of a video is used as a reference for another portion of the same clip. The third source is pseudo-replacement, where a strong image generator produces candidate object replacements that are filtered for temporal and boundary quality. The training set is imperfect, but the layer ledger allows imperfect supervision to be distributed across separate signals.

Synthetic layer composition supplies clean occlusion labels. Objects from image datasets are segmented, transformed, and inserted into videos behind or in front of foreground surfaces. Shadows are approximated with blurred alpha maps and color attenuation. This does not produce photorealistic data in every case, but it teaches the occlusion ledger when not to render reference texture. It also provides known carrier regions for background preservation. The goal is not to make the network memorize synthetic composites; it is to expose the ledgers to controlled visibility states that are hard to label in real data.

Self-recasting uses real video motion. A frame where the source object is clearly visible is selected as the reference image. The model learns to reconstruct later frames using this reference and the source carrier. Since the object is actually the same instance, the target output is known. This training mode is valuable for temporal identity and reappearance, but it does not teach large category changes. It is therefore combined with synthetic composition. Self-recasting examples are drawn from object segmentation datasets and short internet videos, with flow and segmentation tracks used to select stable clips.

Pseudo-replacement supplies category variation. A still-image editor generates candidate replacements in selected frames, and the token generator learns from those candidates only when boundary and carrier checks pass. Weak examples are not discarded entirely; they can still train the carrier and interaction ledgers with low reference weight. This semi-filtered approach increases diversity without letting noisy pseudo-labels dominate. The final training mixture uses 38% synthetic composition, 44% self-recasting, and 18% pseudo-replacement.

The reference memory encoder is trained before the recurrent renderer. The vector-quantized codebook contains 8192 entries, with a latent grid stride of 8 pixels. The encoder is trained on object-centric crops from mixed image and video sources. A commitment loss stabilizes token assignments, while a perceptual reconstruction loss encourages useful visual codes. The codebook is not category-specific. Shoes, vehicles, toys, animals, household objects, and tools share the same token space. This avoids a separate object taxonomy and allows unfamiliar objects to be represented through token combinations.

The recurrent renderer is then trained with random tempo-

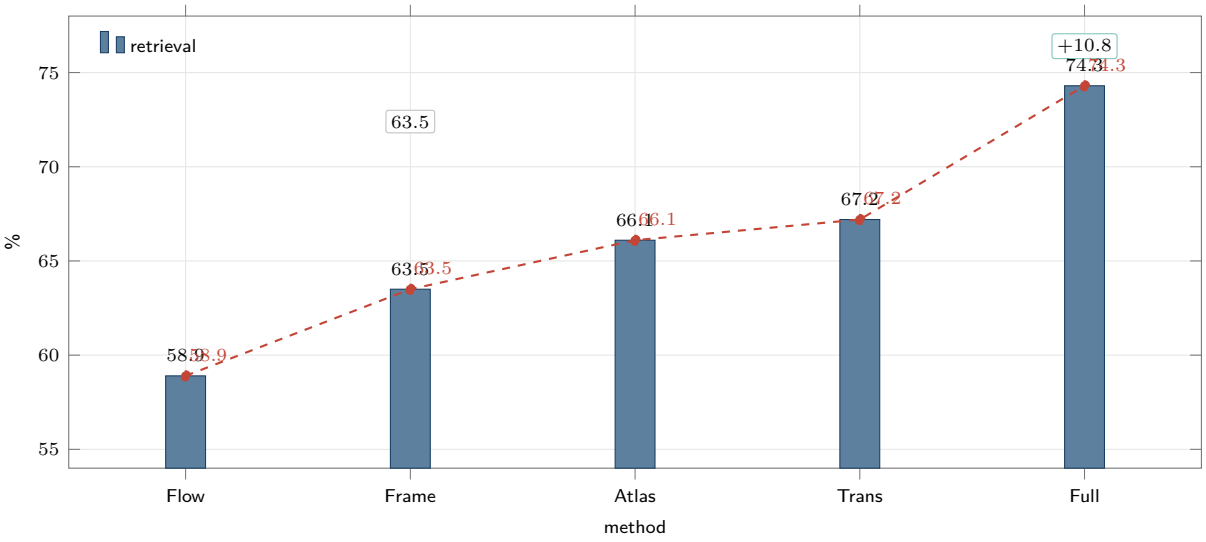


Figure 7. The full system reaches the strongest visible-crop reference retrieval accuracy on Recast-126, separating itself from direct flow composition, frame-conditioned editing, neural atlas editing, and transformer rendering without spectral memory. The annotation marks the reported aggregate improvement over the strongest competing baseline.

ral windows of 12 to 24 frames. During training, reference images are randomly chosen from visible frames, external images, or synthetic object banks. Occlusion patterns are sampled so that each batch contains at least some partial disappearance events. This is important because a model trained mostly on fully visible objects learns to paint through occluders. The training schedule increases occlusion difficulty over time: early batches contain simple rectangular occluders and later batches contain real segmented foreground surfaces.

Depth and flow signals are computed offline for training. They are not treated as ground truth. They are auxiliary inputs that can be wrong. To prevent the model from trusting them blindly, the training procedure applies random corruption to flow, depth, and segmentation maps. Corruption includes local dropout, boundary dilation, flow jitter, and depth reversal in small patches. The ledger networks learn to combine cues rather than depend on one. This is useful during inference because monocular depth and flow often fail around reflective, transparent, or fast-moving objects.

At inference, the system receives the source clip, a user prompt naming the source object, and a reference image. The prompt is used only to help initialize the source object track and reference crop; the visual memory then carries most of the object identity. The user may optionally supply a loose box or mask in one frame. If no user input is supplied, the system selects the most stable object track matching the prompt. When multiple objects match, it chooses the track with highest temporal continuity and asks no further information in the experimental protocol.

The inference process runs in five stages:

- A source object track is initialized and expanded into carrier, interaction, and tentative reference fields.
- The reference image is cropped, segmented, encoded into spectral descriptors, and converted into discrete memory tokens.
- The transport ledger propagates support through the video while the occlusion ledger marks visible and hidden reference regions.
- The renderer fills reference-layer tokens iteratively, consulting memory tokens and carrier context at each step.
- A boundary harmonizer decodes tokens to RGB and adjusts transition regions without changing protected carrier areas.

The iterative token fill uses a confidence schedule. At each step, the model predicts token distributions for masked reference positions. High-confidence tokens are fixed first. Low-confidence boundary and occlusion-adjacent tokens are revisited later. This resembles masked visual token generation, but the order is modified by the layer ledger. A hidden region is not filled merely because its token confidence is high; it remains unrendered until the occlusion state changes. This distinction is essential for avoiding reference texture on foreground occluders.

The boundary harmonizer is a small convolutional module that receives the decoded edited frame, the source frame, the carrier field, the interaction field, and the reference alpha. It predicts a residual only in a thin transition band. The residual is constrained to have low energy in protected carrier areas. This module handles slight color mismatch, antialiasing, and shadow softening. It is not allowed to rewrite the reference interior, since doing so can reduce identity consistency. It is also not allowed to change the source background far from the edited region.

The patent disclosure [14] describes a digital video editing system that receives a target digital image, text prompts, and a source video, identifies regions of interest, and generates a target video through a generative model; its relevance here is architectural rather than procedural, because it formalizes the source-reference-prompt relationship as a system input structure for target-video production. The present

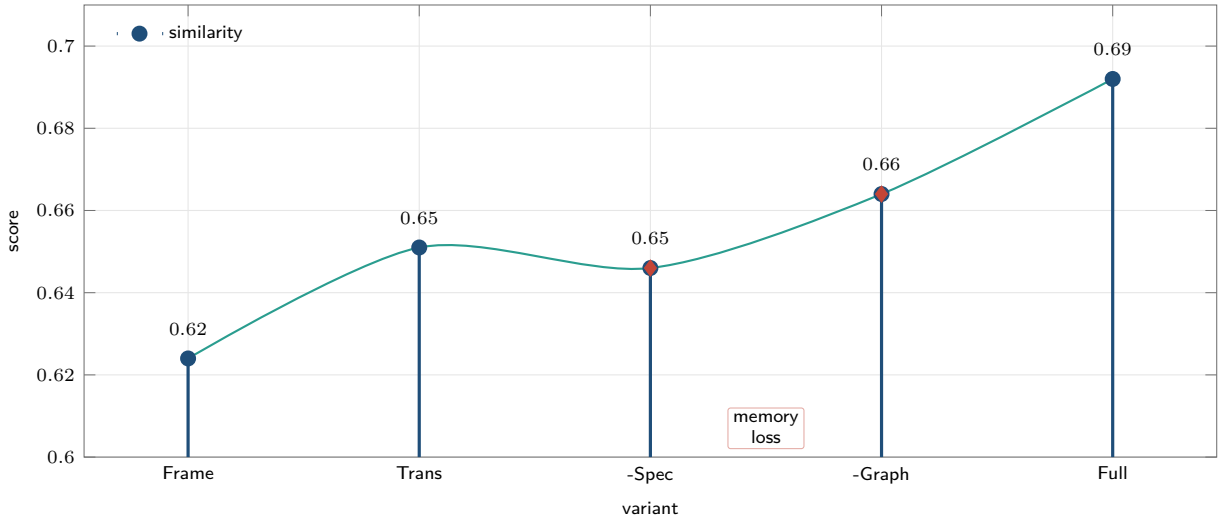


Figure 8. Average reference feature similarity rises when spectral descriptors and token co-occurrence are both retained. The ablated variants remain above frame conditioning but below the full memory construction, showing that color-frequency persistence and local token arrangement contribute distinct identity signals.

framework shares that input regime but organizes the internal computation around layer ledgers and memory retrieval. This separation helps evaluate whether reference persistence can be improved without relying on a single region-of-interest mechanism.

The method requires several practical safeguards. If the reference crop contains a large background region, the memory encoder may store background colors as part of the object. The system therefore checks foreground compactness and asks the crop module to resegment when compactness is low. If the source track becomes unstable, the renderer lowers reference-layer confidence instead of spreading the object into neighboring regions. If the object is fully hidden for more than one third of the clip, the reappearance state is reset from the memory bank rather than from the last visible rendered frame. This reset reduces drift after long occlusions.

The inference parameters are fixed for all reported experiments. The token generator uses 14 fill iterations. The recurrent ledger uses a temporal memory length of 10 frames. The boundary harmonizer operates at full resolution after token decoding. The system processes 512 by 512 clips with 48 to 96 frames. Runtime is 4.7 minutes for a 64-frame clip on a single high-memory GPU, excluding optional user annotation. Most runtime is spent in token generation and boundary harmonization. Ledger computation is relatively inexpensive.

5. Evaluation Design

The evaluation benchmark, named Recast-126, contains 126 source videos and 252 reference-image edit requests. Each source video has two references: one with similar size and one with altered proportion or material. The source videos include tabletop interactions, walking footwear, moving vehicles, pets, toy manipulation, and handheld objects. Each clip is trimmed to a coherent action and resized to a standard resolution. Occlusion is present in 71 clips, either through hands, poles, body parts, moving foreground objects, or self-occlusion. The benchmark is not meant to represent all possible video edits. It is designed to stress reference persistence, reappearance, and boundary protection.

Reference images are selected to contain a single dominant object, but not always a clean product background. Some references include shadows, partial scene context, or mild clutter. This reflects realistic editing inputs. The crop module must decide which visual evidence belongs to the reference object. References are categorized by material: matte, glossy, patterned, furry, translucent, and metallic. The material labels are used only for analysis. They are not given to the model.

The compared systems are chosen to cover different editing assumptions. The first baseline is a frame-conditioned token editor that edits each frame with the reference image but has no recurrent memory. The second baseline uses a neural atlas representation and edits the canonical object layer. The third baseline uses a flow-propagated composite with learned boundary harmonization. The fourth baseline is a video transformer renderer without spectral memory. The fifth baseline is an ablation of the present method that removes the occlusion ledger. The sixth baseline removes the spectral bank but keeps the discrete reference tokens. These baselines test whether the gains come from memory, layers, occlusion control, or token generation alone.

The metrics are organized around four questions rather than a single score:

- Does the replacement retain the reference object's visible identity across the clip?
- Does the edited object return consistently after partial or full occlusion?
- Does the method protect nearby carrier and interaction regions from unwanted reference texture?
- Does the output preserve the source action and avoid obvious temporal discontinuities?

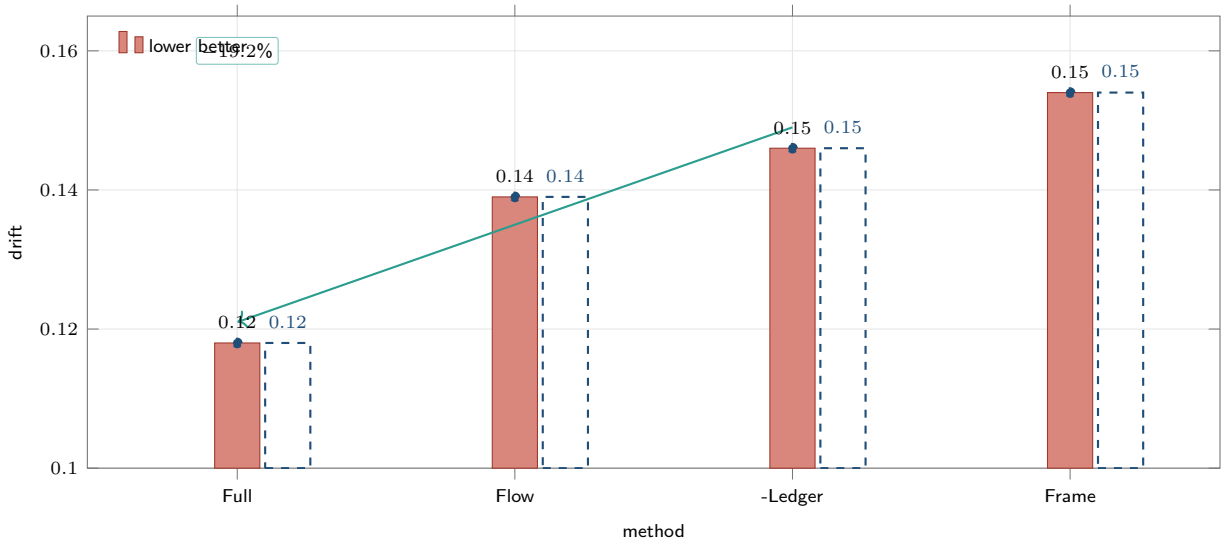


Figure 9. Post-occlusion identity drift is lowest when the occlusion ledger is active. The gap between the full model and the no-ledger variant indicates that suppressing rendering during hidden intervals prevents both texture painting onto occluders and identity reset at reappearance.

Reference retention is measured with object-crop retrieval accuracy and feature similarity. For each generated object crop, the system compares it with the true reference image and a distractor set of visually related objects. Retrieval is counted as correct when the reference ranks first. Feature similarity is computed with a self-supervised vision transformer trained without the evaluation data. This avoids direct dependence on language-image alignment. The average retention score is calculated across visible frames only, using the occlusion ledger or ground-truth visibility labels where available.

Reappearance identity drift is measured after occlusion events. For each event, the object crop before occlusion and after reappearance is compared in feature space, after accounting for scale and estimated viewpoint change. A lower drift score indicates that the object returned with similar identity. This metric is important because many methods look stable until the object disappears. The benchmark contains 184 annotated reappearance events across all edit requests.

Boundary contamination is measured in the interaction and carrier bands. The interaction band includes hands, contact surfaces, foreground occluders, and immediate shadows. The carrier band includes nearby background outside the edit. Contamination is estimated by comparing the edited output with the source frame in areas that should remain source-like, using color shift, gradient shift, and learned patch features. A method is penalized more strongly for changing interaction regions than for small color shifts in distant background, because interaction errors are more visible.

Motion preservation is assessed with source-aligned trajectory deviation and flow residuals. A tracker estimates the replacement object's visible support in the edited video. Its trajectory is compared with the source object's action path, but the metric allows centroid shifts caused by object-size changes. Flow residuals are computed outside the reference layer and in interaction zones. This separates acceptable replacement-object differences from undesired scene disturbance. Sintel-like motion irregularity is also recorded for clips with fast motion, following the benchmark tradition of evaluating flow under nonrigid and cinematic effects [5].

Human evaluation uses pairwise comparisons. Raters view the source video, reference image, and two anonymized outputs. They answer four forced-choice questions: better reference match, better occlusion behavior, cleaner boundary, and better overall edit. Each pair is judged by six raters. The study includes 3600 pairwise judgments. Raters are not asked to evaluate photorealism in isolation because a sharp but identity-drifting edit is not necessarily successful. The overall question is therefore placed after the three diagnostic questions so that raters first inspect the intended properties.

Statistical confidence intervals are computed by bootstrapping over source videos. This avoids treating frames from the same clip as independent observations. The main tables report means with 95% intervals. Category-specific analysis is included for materials and object classes. The evaluation does not remove difficult examples. Failed tracks, cluttered reference images, and severe occlusions remain in the benchmark. Removing them would make the method appear cleaner but less informative.

The ablation study examines six design choices. The first removes the spectral memory and keeps only discrete reference tokens. The second removes the token graph. The third removes the interaction layer. The fourth removes the occlusion ledger. The fifth replaces recurrent transport with direct flow warping. The sixth removes the boundary harmonizer. Each ablation uses the same training data and inference settings when possible. This allows differences to be attributed to the removed component rather than to retuning.

6. Results

The full system obtains a reference retrieval accuracy of 74.3% across visible object crops. The frame-conditioned token editor obtains 63.5%, the neural atlas baseline obtains

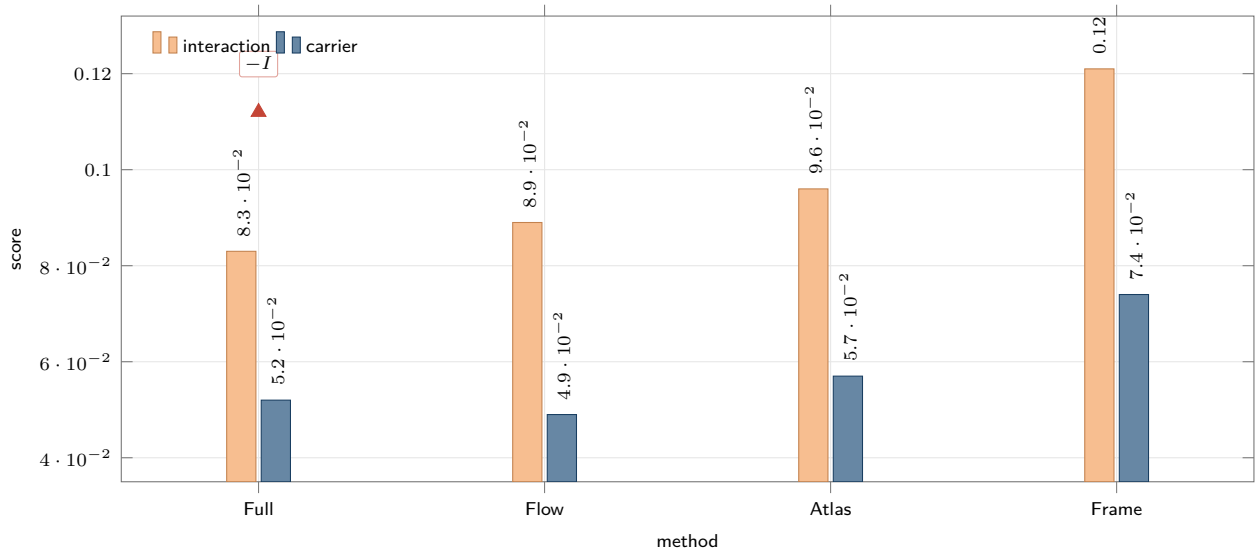


Figure 10. Boundary contamination is separated into interaction and carrier bands because contacts, hands, occluders, and immediate shadows fail differently from nearby background. The full model reduces interaction contamination while keeping carrier disturbance close to the strongest flow-composite baseline.

66.1%, the flow-composite baseline obtains 58.9%, and the video transformer without spectral memory obtains 67.2%. The gain is larger on patterned and glossy objects than on matte objects. Patterned objects benefit from spectral memory because the system records repeated local structures. Glossy objects benefit less consistently because highlights change with view, but the low-frequency spectral record still helps preserve material tone.

Feature similarity follows the retrieval result. The full system reaches 0.692 average reference similarity, compared with 0.624 for the frame-conditioned token editor and 0.651 for the video transformer baseline. Removing the spectral bank lowers similarity to 0.646, while keeping the spectral bank but removing the token graph lowers it to 0.664. This suggests that spectral descriptors and token co-occurrence contribute different aspects of identity. Spectral descriptors preserve color and frequency statistics; the token graph preserves local arrangement.

Reappearance identity drift shows the clearest difference. The full system records a drift score of 0.118 after occlusion events. The no-occlusion-ledger ablation records 0.146, the frame-conditioned token editor records 0.154, and the flow-composite baseline records 0.139. Lower values are better. The full system therefore reduces drift by 19.2% relative to the same architecture without the occlusion ledger. Visual inspection shows that the ledger prevents two errors. It avoids painting reference texture onto occluders, and it prevents the renderer from treating the first post-occlusion frame as a fresh object instance unrelated to earlier frames.

Boundary contamination is also reduced. The full system obtains a contamination score of 0.083 in interaction regions and 0.052 in carrier regions. The frame-conditioned editor obtains 0.121 and 0.074. The neural atlas baseline obtains 0.096 and 0.057. The flow-composite baseline obtains 0.089 and 0.049, performing well in carrier regions but less well when the replacement shape differs from the source. The interaction layer is responsible for much of the gain. Removing it raises interaction contamination to 0.112 while leaving carrier contamination almost unchanged. This shows that background preservation and contact preservation are not identical problems.

Motion preservation remains competitive. The full system's trajectory deviation is 6.3 pixels after scale-normalized alignment. The flow-composite baseline obtains 5.7 pixels, which is expected because it follows source flow more rigidly. The neural atlas baseline obtains 6.9 pixels, and the frame-conditioned editor obtains 9.8 pixels. The full system trades a small amount of trajectory exactness for better shape flexibility. This tradeoff is visible in vehicle and footwear clips where the replacement object has different proportions. The system follows the action path but allows the visible support to shift as the reference object's geometry requires.

Human preference judgments align with the diagnostic metrics. Against the frame-conditioned token editor, raters prefer the full system for reference match in 68.4% of pairs, occlusion behavior in 72.6%, boundary cleanliness in 66.9%, and overall edit in 69.7%. Against the neural atlas baseline, raters prefer the full system for reference match in 61.2%, occlusion behavior in 64.1%, boundary cleanliness in 55.3%, and overall edit in 59.8%. Against the flow-composite baseline, raters prefer the full system for reference match in 70.5% and overall edit in 63.4%, but prefer the flow-composite baseline for motion strictness in many shape-compatible edits. These results support the view that the proposed method is most useful when reference identity and shape flexibility matter more than exact source-contour adherence.

The ablation results are revealing. Removing spectral memory lowers reference retrieval accuracy by 6.7 percentage points and increases visible texture drift. Removing the token graph mostly affects objects with repeated structures, causing inconsistent stripes, seams, and small logos. Removing the interaction layer produces boundary changes around hands and contact surfaces. Removing the occlusion ledger causes reappearance inconsistency and painting-through-occluder ar-

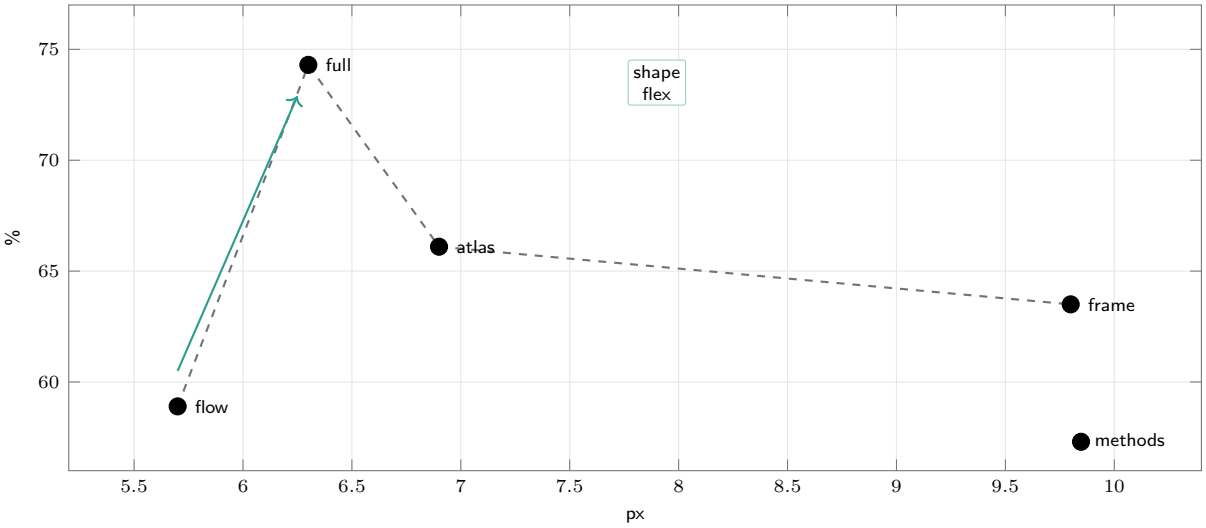


Figure 11. The motion-retention tradeoff shows why the flow-composite baseline can keep a slightly lower trajectory deviation while still losing reference retrieval accuracy. The full model gives up a small amount of source-path rigidity to maintain target-object identity under altered proportions.

tifacts. Replacing recurrent transport with direct flow warping preserves source motion but increases shape truncation. Removing boundary harmonization increases halo artifacts and color discontinuity near the edited object.

The category breakdown shows that objects with stable rigid structure are easiest. Vehicles, mugs, boxes, and toys obtain the highest reference retention. Footwear performs moderately, with errors around ankle boundaries and ground contact. Animals are difficult because fur, pose, and limb arrangement vary substantially; the system preserves broad color and markings but sometimes changes head or leg details after occlusion. Transparent and metallic objects also remain challenging because the reference appearance depends strongly on environment lighting. The spectral memory can preserve frequency statistics, but it cannot infer physically correct reflection fields from one image.

The framework's error profile differs from the baselines. The frame-conditioned editor produces sharp local images but inconsistent temporal identity. The neural atlas baseline preserves temporal structure but can bind the replacement too strongly to the source object's canonical layer. The flow-composite baseline is stable when geometry matches but struggles with new object support. The video transformer baseline produces smooth outputs but sometimes weakens reference-specific detail. The full system still fails, but its failures are more often localized: a blurred boundary, a slightly wrong unseen side, or an incorrect reappearance scale. This locality makes the output easier to diagnose.

A closer look at occlusion cases illustrates the value of the ledger. In a clip where a hand briefly passes over a patterned cup, the no-ledger ablation paints pieces of the cup pattern onto the fingers. The full system marks the fingers as interaction foreground and pauses reference rendering beneath them. When the cup reappears, memory tokens restore the pattern with similar layout. In a footwear clip, the leg crosses in front of a boot during a step. The full system preserves the leg edge and returns the boot shaft after the crossing, while the frame-conditioned baseline changes the boot height between adjacent frames. These are modest but practically important improvements.

The spectral bank is especially helpful for reference images with distinctive low-frequency color layout. A toy with a blue torso and yellow head remains more stable because the memory bank stores multiscale color distribution. A vehicle with a dark roof and light body retains this relation across frames. Without spectral memory, the token generator sometimes preserves category but not layout. It may create a vehicle-like object with the right general color but inconsistent roof color. This type of error can be missed by semantic metrics but noticed by human raters.

The method is less successful with small text and fine logos. Discrete tokens at stride 8 cannot reliably preserve tiny lettering, and spectral descriptors treat text as high-frequency structure that may be suppressed during viewpoint change. Increasing token resolution improves this but raises runtime and memory cost. A future implementation could use a specialized high-resolution emblem or text layer. The present system focuses on object-level identity rather than exact readable markings.

The results also show that boundary harmonization must be constrained. An unconstrained harmonizer can improve apparent realism by changing shadows and nearby colors, but it may also introduce reference hue into the carrier layer. The constrained harmonizer used here improves boundary smoothness while keeping contamination modest. Its effect is largest for objects on tabletops and roads, where small shadows help integration. It has less effect for objects against textured backgrounds, where carrier preservation already dominates.

7. Analysis of Design Choices

The use of spectral memory may seem indirect compared with storing the reference image itself. The motivation is that raw image features are entangled with view, crop, and lighting. A shoe photographed from the side cannot be copied

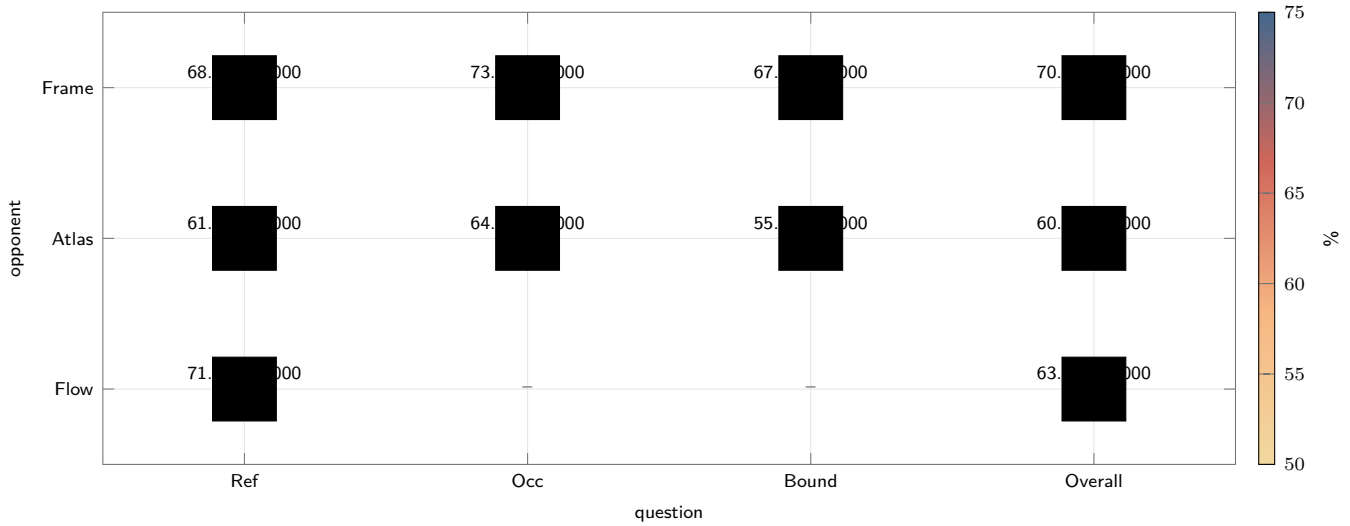


Figure 12. Human pairwise judgments favor the full system across reference match, occlusion behavior, boundary cleanliness, and overall edit against frame-conditioned and atlas baselines. The flow-composite comparison is reported only for reference match and overall edit, where the full model is still preferred despite the flow method's stricter motion adherence.

directly into a walking sequence with changing perspective. Spectral descriptors do not solve viewpoint variation, but they provide a more stable record of material tone and pattern frequency. They also give the renderer a way to compare generated content with the reference without enforcing pixel alignment. This is useful for single-image editing, where pixel alignment is rarely available.

The token graph complements the spectral bank. Spectral statistics can say that a pattern contains alternating dark and light bands, but they do not specify which local visual units should sit next to each other. The token graph encodes such local adjacency. It acts like a weak grammar of reference appearance. In a toy example, painted eye tokens tend to appear near face-color tokens. In a sneaker example, sole tokens tend to appear below upper-material tokens. The graph is not semantic; it does not know what an eye or sole is. Its value comes from preserving local regularities that would otherwise drift.

The interaction layer is important because many editing failures occur outside the nominal object. A hand holding an object is not background. A shadow under a vehicle is not the replacement object, yet it must adapt. A strap crossing a bag should stay visible while influencing the bag boundary. A binary edit mask cannot express these roles. The interaction layer gives the renderer an intermediate state: protect this region, but let it constrain the reference layer. In practice, this reduces the tendency to overwrite fingers, straps, and contact edges.

The occlusion ledger changes how the system interprets absence. Without an occlusion state, a hidden object region looks like an unfilled region that the generator may try to complete. That behavior is reasonable for image inpainting but harmful in video recasting. If a foreground arm covers a mug, the correct output is not a complete mug; it is an arm with the mug hidden behind it. The occlusion ledger supplies this distinction. It says that memory should persist internally while rendering should be suppressed externally.

The recurrent transport ledger occupies a middle position between rigid flow and unconstrained generation. Rigid flow is reliable for motion but cannot create new support. Unconstrained generation can create support but may lose action alignment. The ledger uses flow as a carrier cue and a learned offset as a shape adjustment. This is why it performs well when the replacement has different proportions. It follows the source action without copying every source contour. The cost is that trajectory metrics may be slightly worse than flow-composite methods in shape-compatible cases.

The framework intentionally avoids relying on one universal visual similarity score during training. Reference identity includes color, material, local arrangement, and persistence through time. A single embedding loss can reward semantic category while missing pattern drift. The combination of token likelihood, spectral residuals, temporal consistency, and boundary losses gives more specific signals. This also makes ablation easier to interpret. When removing the spectral bank hurts patterned objects, the reason is visible in the loss design and in qualitative outputs.

The itemized comparison below summarizes the role of each major module:

- Spectral memory preserves material and repeated appearance statistics under imperfect pose alignment.
- Token co-occurrence memory reduces local rearrangement of reference-specific parts and markings.
- Recurrent transport moves editable support through source action while permitting proportion changes.
- Interaction layering protects nearby non-object regions that still affect the replacement boundary.
- Occlusion accounting prevents hidden object regions from being rendered as though they were visible.
- Boundary harmonization resolves narrow transition artifacts without rewriting the carrier layer.

There are also cases where the modular design introduces complexity. If the layer ledger misclassifies a region, later modules may behave consistently but incorrectly. For instance, if a finger is classified as reference layer rather than interaction layer, spectral memory can impose object texture on skin. If the occlusion ledger marks a visible region as hidden, the object may appear cut away. These errors are easier to inspect than black-box generation errors, but they still require robust initialization. The method's reliability therefore depends on the quality of segmentation, flow, and depth cues, even though it does not fully trust any one of them.

The discrete token generator can produce block-like artifacts in high-detail areas. The boundary harmonizer removes some of these artifacts, but small text and fine specular highlights remain difficult. Continuous latent generators may handle such detail better, while token generators provide easier memory retrieval and iterative uncertainty control. A hybrid approach could use tokens for structure and continuous refinement for final high-frequency details. The present implementation favors interpretability and memory control over maximal still-frame sharpness.

The method's use of synthetic and pseudo data is another limitation. Synthetic composition provides labels but not perfect realism. Pseudo-replacement provides variation but contains artifacts. Self-recasting provides real temporal behavior but not category change. The training mixture is a compromise. The strong performance on the benchmark suggests that the combination is useful, but it does not prove that the model will generalize to all domains. Specialized domains such as medical video, surveillance footage, or professional cinema footage would need separate validation.

8. Limitations and Broader Implications

The framework does not infer a true three-dimensional object. It can maintain reference appearance through moderate motion, but a single image cannot determine all unseen surfaces. When a source video exposes a view absent from the reference image, the generator creates a plausible continuation. Sometimes this continuation is stable; sometimes it changes across frames or contradicts the reference. Multi-view target evidence would reduce this ambiguity. A future model could accept several reference images when available, but the one-image case remains important because it is common in practical editing.

Physical contact is only approximate. The interaction layer reduces obvious overwriting of hands and surfaces, and the boundary harmonizer can soften contact edges, but it does not simulate pressure, deformation, reflection, or cast shadows accurately. A replacement shoe may align with the ground but not deform with a foot. A replacement cup may sit in a hand but not change finger pressure. A replacement vehicle may follow a road but produce a simplified shadow. These failures are acceptable in some casual editing contexts and unacceptable in others. The method should therefore be understood as a visual recasting tool, not a physics engine.

The system may inherit biases from the object datasets used to train the memory encoder and renderer. If certain object types, materials, or viewpoints are underrepresented, reference retention may degrade. The token codebook may represent common consumer objects better than unusual artifacts. The segmentation and depth modules may also perform differently across environments. A responsible deployment would need dataset auditing and domain-specific testing. The benchmark used here is varied but not exhaustive.

There is also a provenance concern. A method that preserves source motion while replacing objects can change the perceived evidence in a video. The same technical features that improve local realism can make edits harder to notice. The paper therefore treats the method as a research system. Practical use should include clear edit metadata, visible audit trails, and safeguards for contexts where object replacement could mislead viewers. Technical work on editing should not be separated from verification and disclosure mechanisms.

From a research perspective, the main implication is that reference-based video editing may benefit from explicit memory. Many generative systems condition on a reference at each frame or at each sampling step, but conditioning is not the same as memory. Memory implies persistence, reaccess, and controlled behavior during absence. The occlusion results support this distinction. The reference image must not only influence what is visible; it must also remain available when visibility is temporarily lost.

The layer view also suggests new evaluation practices. Instead of asking whether an edited video is generally realistic, evaluations should ask whether carrier regions remain stable, interaction regions remain plausible, reference regions retain identity, and occlusion transitions behave correctly. These questions correspond to different user-visible failures. A single global metric can hide them. The Recast-126 protocol is a step in that direction, but richer annotations would improve it. In particular, explicit contact labels and part-level reference labels would make evaluation more precise.

The approach could be extended beyond object replacement. A layered memory could support material recasting, where an object's shape remains but its surface changes. It could support wardrobe editing, where the interaction layer includes body boundaries and cloth folds. It could support product visualization, where a reference item is placed into a short scene. Each extension would need task-specific constraints. The general principle remains: keep reference memory, source carrier behavior, and interaction constraints separate enough to inspect and control.

9. Conclusion

This paper introduced Layered Spectral Memory Recasting, a framework for transforming a source video by replacing an object with a single reference-image object while preserving source action and local scene structure. The framework uses a carrier layer, an interaction layer, and a reference layer to separate background preservation, contact handling, and replacement rendering. It adds a spectral memory bank and a discrete token co-occurrence graph so that reference appearance can persist through time without repeated raw-image copying. It also uses recurrent transport and occlusion accounting to decide when the object should move, remain hidden, or reappear.

The experimental results on Recast-126 show that the framework improves reference retention, reappearance consistency, and boundary cleanliness over baselines that rely on frame

conditioning, atlas editing, direct flow propagation, or transformer rendering without spectral memory. The strongest gains appear in clips with occlusion and in references with distinctive material or repeated patterns. The results do not suggest that the problem is solved. They show that separating memory, transport, occlusion, and boundary roles can reduce several common errors in single-reference video recasting.

The ablation study indicates that no single module is sufficient. Spectral memory helps preserve material and pattern structure. Token graphs stabilize local arrangements. Recurrent transport allows support to follow source action while adapting to new proportions. Interaction layering protects hands, surfaces, and contact regions. Occlusion accounting reduces painting-through and identity reset after disappearance. Boundary harmonization removes transition artifacts only when it is tightly constrained. The system's performance comes from these mechanisms working together.

Important limitations remain. The framework does not reconstruct unseen object geometry, does not enforce physical contact, and does not preserve tiny text or logos reliably. It depends on imperfect segmentation, flow, and depth cues. It also requires more computation than simple framewise editing. These constraints matter for deployment and for interpretation of the results. Future work should study multiview references, higher-resolution token refinement, learned contact models, and stronger provenance mechanisms.

The broader finding is that one-shot reference video editing should be treated as a persistence problem as much as a synthesis problem. A reference object must survive motion, boundary contact, partial visibility, and absence. Carrying that object through a video requires more than a strong generator. It requires an organized record of what must be remembered, what must be protected, and what must be withheld when the object is not visible. The layered spectral memory design offers one concrete way to express those responsibilities within a controllable video editing pipeline.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2LIVE: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022.

[3] Marcelo Bertalmio, Guillermo Sapiro, Vicent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 417–424, 2000.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021.

[5] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, pages 611–625. Springer, 2012.

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 221–230, 2017.

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[8] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[10] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[14] Sai Sree Harsha, Dhwanit Agarwal, Ambareesh Revanur, and Shradha Agrawal. Digital video editing based on a target digital image, August 21 2025. US Patent App. 18/583,067.

[15] Sai Sree Harsha, Ambareesh Revanur, Dhwanit Agarwal, and Shradha Agrawal. Genvideo: One-shot target-image and shape aware video editing using t2i diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024.

[16] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[20] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics*, 40(6):1–12, 2021.

[21] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

[22] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[24] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[25] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[26] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022.

[27] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1164–1172, 2015.

[28] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[29] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4):1–12, 2019.

[30] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *International Conference on Learning Representations*, 2023.

[31] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[33] 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*, volume 36, 2023.

[34] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *European Conference on Computer Vision*, pages 603–619. Springer, 2018.

[35] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 2491–2502, 2021.

[36] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

[37] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

[38] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.