Critical Review of "AI-Powered Video Editing: Motion-Aware Object Replacement Using Generative Models" (Revised for Literature Positioning and Mathematical Clarity)

Stepan Goyunyan*, Narek Meliksetyan*, Arthur Tsaturyan†, Elizaveta Labazanova‡, Zangir Iklassov§, Jorge Alejandro Amad
*Physmath School After Artashes Shahinyan, Yerevan, Armenia
†Evrika STEM Specialized School, Vanadzor, Armenia
‡Moscow, Russia
§Mohammad Bin Zayed University of Artificial Intelligence (MBZUAI), UAE

***Abstract*—This document critically reviews Aghababyan's 2025 master's thesis on motion-aware object replacement in videos using a modular pipeline (detection/segmentation, motion handling, and diffusion-based inpainting). The review summarizes the reported approach, identifies strengths and limitations grounded in the thesis text, and positions the work relative to task-adjacent and task-specific literature on temporally consistent video editing and video inpainting. To address feedback on mathematical rigor, the review formalizes the thesis motion modules using standard notation and derives several limitations that follow directly from the stated transformations. External references are used only for contextual positioning and for standard evaluation practice; they are not used to make empirical claims about the thesis results.**

***Index Terms*—video editing, object replacement, video inpainting, diffusion models, temporal consistency, optical flow**

Primary target document: Satine Aghababyan, *AI-Powered Video Editing: Motion-Aware Object Replacement Using Generative Models*, M.S. thesis, American University of Armenia, 2025 [1].

Verification policy: Statements about the thesis are limited to what is explicitly described, implemented, or reported in the thesis text and figures. Statements about standard practice or alternative approaches are supported by external literature and are presented as context rather than as claims about the thesis' empirical outcomes.

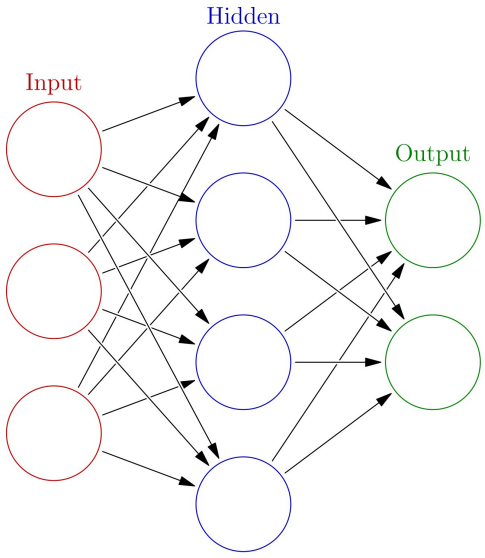

Fig. 1. Generic schematic of a feed-forward neural network. The thesis pipeline uses multiple learned modules (segmentation/detection, optical flow, diffusion denoising, and conditioning adapters), each implemented as a neural network with task-specific architecture.

I. Overview of the Thesis Approach

The thesis proposes a modular pipeline for replacing a selected object throughout a video while aiming to preserve realism and temporal consistency.

The pipeline combines object detection and segmentation (reported as YOLOv8), motion handling via either 2D geometric trajectory propagation or optical-flow-based warping (RAFT) [2], and diffusion-based inpainting in the latent space of Stable Diffusion style backbones [3]. To reduce frame-to-frame variability in diffusion outputs, the thesis evaluates controlled inpainting mechanisms including fixed latent initialization, ControlNet structural conditioning (notably monocular depth maps) [4], and IP-Adapter image prompting [5]. The thesis also reports negative results for latent-space warping.

II. Strengths

- The thesis correctly identifies the core instability of frame-wise diffusion inpainting for videos, namely stochastic variation that breaks temporal coherence.
- The thesis compares multiple practical consistency mechanisms (trajectory transfer, optical flow warping, Control-

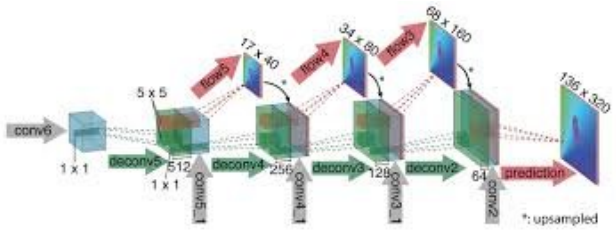

Fig. 2. Illustrative U-Net style encoder-decoder architecture, commonly used as the denoising backbone inside diffusion pipelines (including latent diffusion variants). Included to contextualize the diffusion-based inpainting component described in the thesis.

Fig. 3. Example of articulated human motion across frames (non-rigid motion). Such motion changes internal configuration in ways a single global similarity transform cannot represent, consistent with the derived limitation following Eq. (5).

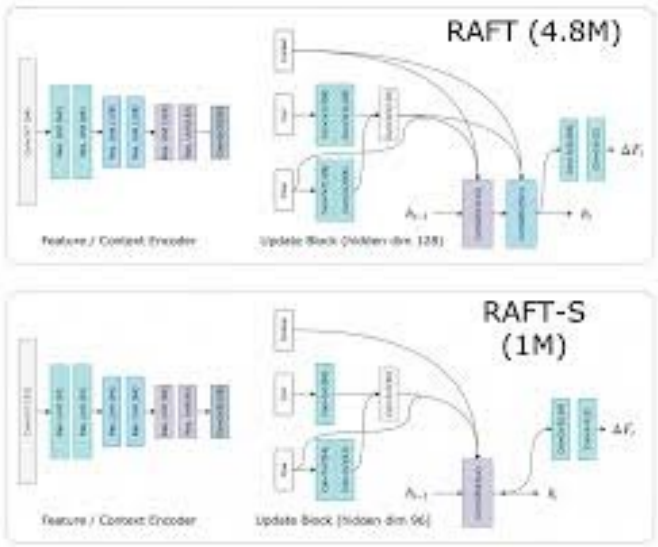

Fig. 4. High-level schematic of the RAFT optical flow network (and a smaller RAFT-S variant). Conceptually, RAFT extracts features, builds an all-pairs correlation volume, and iteratively refines the flow estimate with a recurrent update operator [2].

Net depth conditioning, IP-Adapter, fixed latent initialization), including negative results for latent-space warping.
- The thesis motivates reuse of pretrained components rather than training a large video generator from scratch.

III. Mathematical Clarification of Key Mechanisms (Derived from Thesis Descriptions)

This section does not attribute new algorithms to the thesis. It expresses thesis-described steps in standard notation to remove ambiguity and to make several limitations precise.

A. Notation

Let a video be a sequence of frames $\{I_t\}_{t=1}^{T}$ with pixel domain $\Omega \subset \mathbb{R}^2$. Let $M_t \subset \Omega$ denote the binary mask of the target object in frame t obtained by per-frame segmentation. Let the edited output frame be $\hat{I}_t$.

B. 2D Geometric Trajectory Propagation (Similarity Transform)

The trajectory approach computes translation, scaling, and rotation between consecutive frames and applies them to the replacement object. A standard formulation is the similarity transform

$$T_t(x) = s_t R(\theta_t) x + b_t, \tag{1}$$

where $s_t > 0$ is a scale, $R(\theta_t)$ is a 2×2 rotation matrix, and b_t is a translation. One mask-derived translation is centroid displacement:

$$c_t = \frac{1}{|M_t|} \sum_{x \in M_t} x, \qquad b_t = c_{t+1} - c_t. \tag{2}$$

If a "diameter" proxy is derived from the maximum radius from the centroid, one natural proxy is

$$d_t = 2 \max_{x \in M_t} \|x - c_t\|_2. \tag{3}$$

The thesis defines a size ratio $r = d_{t+1}^2 / d_t^2$, corresponding to an area ratio under a circular proxy. The corresponding linear scale factor is

$$s_t = \frac{d_{t+1}}{d_t} = \sqrt{r}. \tag{4}$$

Derived limitation (non-rigid motion). Any transform of the form (1) preserves pairwise distances up to a global factor:

$$\|T_t(x_i) - T_t(x_j)\|_2 = s_t \|x_i - x_j\|_2 \qquad \forall x_i, x_j. \tag{5}$$

Therefore, if the target object undergoes non-rigid deformation where distances between at least one pair of material points change non-uniformly over time, no single similarity transform can match that deformation. This formalizes why translation/rotation/scale tracking cannot represent articulated motion (for example, humans) even if centroid tracking is accurate.

C. Optical-Flow Warping

Let $F_{t+1 \to t}(x) = (u_{t+1 \to t}(x), v_{t+1 \to t}(x))$ denote a backward dense optical flow field mapping a pixel location x in frame $t+1$ to a corresponding location in frame t [2]. Figure 4 provides a high-level schematic of the RAFT architecture. A standard backward warp of an edited frame $\hat{I}_t$ into frame $t+1$ is

$$W(\hat{I}_t, F_{t+1 \to t})(x) = \hat{I}_t\big(x + F_{t+1 \to t}(x)\big), \tag{6}$$

with interpolation for non-integer coordinates.

D. Temporal Consistency as a Measurable Constraint

A practical way to quantify temporal stability is a warping-consistency error over visible regions. Let $\Omega_{t+1}^{\text{vis}} \subseteq \Omega$ denote

non-occluded pixels in frame $t+1$. A simple measure is

$$E_{\text{warp}}(t) = \frac{1}{|\Omega_{t+1}^{\text{vis}}|} \sum_{x \in \Omega_{t+1}^{\text{vis}}} \left\| \hat{I}_{t+1}(x) - W(\hat{I}_t, F_{t+1\to t})(x) \right\|_1 . \tag{7}$$

This definition is included to make "temporal consistency" operational and to support the evaluation recommendations in Sec. VII. It does not claim that the thesis used this metric.

IV. Limitations and Weaknesses Supported by the Thesis Text

A. Temporal Consistency is Not Enforced End-to-End

Temporal coherence is approached through auxiliary controls around per-frame generation rather than through a unified video-native generative model.

B. Per-Frame Mask Generation Without a Described Temporal Stabilization Procedure

Mask generation is performed per frame via detection and segmentation. The thesis does not describe a dedicated temporal mask stabilization or tracking-based mask propagation step. Consequently, the reviewed thesis does not establish how mask consistency is maintained across time.

C. Motion Handling Limitations

Trajectory propagation. The thesis reports that the trajectory method works primarily for rigid objects, can miss internal articulation (for example wheel spin), and fails for non-rigid objects. Equation (5) makes this limitation explicit.

Optical-flow warping (RAFT). The thesis reports practical artifacts such as blurring and stretching under drastic motion and notes degraded performance on real-content videos, with discussion that RAFT is commonly trained on synthetic datasets [2]. Occlusions and dis-occlusions are discussed and occlusion masks are computed, but the thesis does not fully specify a complete, reproducible frame update rule for how occluded regions are integrated into the edited sequence.

D. Controlled Inpainting Improves Stability but Remains Limited

The thesis states that uncontrolled per-frame diffusion inpainting produces different replacement objects across frames even when using fixed seeds. Controlled mechanisms improve stability, but the thesis reports that full temporal consistency is still not achieved. The thesis also notes a structural limitation of IP-Adapter use: the pipeline cannot provide a mask for the reference image, so correspondence between the masked region and reference features is not explicitly constrained [5]. The thesis further reports that multiple conditions can conflict and misguide generation.

E. Latent-Space Warping is Reported to Fail Quickly

The thesis experiments with warping in latent space using optical flow and reports rapid degradation, including outputs that "fade out" after one or two generations.

F. Proposed Optical-Flow-Matching Optimization is Under-Specified

As future work, the thesis proposes minimizing the difference between optical flow of original frames and optical flow of edited frames via latent optimization and limited training. However, the thesis does not fully specify reproducibility-critical details such as what parameters are trained, objective scope and regularization, stopping criteria, and how tradeoffs are balanced between motion matching and visual quality.

G. Technical Clarity and Reproducibility Gaps

1) **Model description inconsistency:** the methodology reports YOLOv8, while the preliminary description resembles early YOLO formulations, reducing clarity.
2) **Optical flow notation:** optical flow is presented in a way that can be read as depending on a single coordinate rather than full pixel location.
3) **Scale application ambiguity:** the trajectory method defines $r = d_{t+1}^2/d_t^2$ (area ratio under the circular proxy) but does not specify the linear scaling factor used when applying the scaling transform. Equation (4) shows that a linear scale should be $\sqrt{r}$ if the intent is to match the diameter proxy.
4) **Incomplete experimental protocol:** the thesis does not provide a complete, reproducible protocol, including dataset specification, diffusion sampling hyperparameters, prompts and seeds, and compute and runtime reporting.

H. Positioning and Narrative Balance (Structural Limitations of the Write-up)

Two additional limitations concern paper structure rather than algorithm design:

1) **Related work coverage for task positioning:** the thesis places more emphasis on component-level background (diffusion models, optical flow, and tool modules) than on task-specific surveys of video object replacement, video inpainting, and temporally consistent video editing. This makes it harder to interpret the contribution relative to established baselines and typical evaluation protocols in this area. The existence of dedicated text-guided video editing benchmarks and challenge tracks (for example LOVEU TGVE) underscores that task-specific evaluation is available [18], [19].
2) **Preliminaries versus evaluation space:** the preliminaries are written in a tutorial style with extensive background and figures for multiple modules. While these descriptions may be correct, the imbalance reduces space available for systematic evaluation, failure analysis, and reproducibility details.

I. Compute Claims and Resource-Efficiency Reporting

The thesis motivates resource efficiency via reuse of pre-trained evaluation components, but it does not report concrete compute requirements (hardware, runtime per frame, memory

footprint). Given increasing interest in efficient, resource-constrained deployment in adjacent vision-and-robotics settings, it would strengthen the thesis to quantify computational cost and to discuss feasible lightweight variants where applicable [24], [25].

V. Supplementary Qualitative Reproduction

Separately from the thesis text itself, a reviewer-run notebook result was supplied as a qualitative supplement to this review. Figure 5 shows a 12-frame grid from that run, where a small cat in the source clip is progressively replaced by a stylized toy-like character. The sequence is informative because it captures both a partial success and a central failure mode discussed throughout this review. Coarse spatial localization is largely maintained: the edited object remains in approximately the intended ground-plane region across frames. However, object identity is not temporally stable. The generated replacement changes substantially in color, facial structure, proportions, and pose across time, with the sequence drifting from a more cat-like yellow object in earlier edited frames toward a red, rounded robot-like character in later frames. This behavior is qualitatively consistent with the review's main critique that auxiliary controls around frame-wise generation can preserve rough placement while still failing to enforce stable cross-frame appearance.

Because the full notebook settings, prompts, and seeds are not documented inside this manuscript, this figure should be interpreted as qualitative supporting evidence rather than as a controlled benchmark comparison. Its value is diagnostic: it visually illustrates how temporal inconsistency can remain significant even when the edited object stays in approximately the correct scene location.

VI. Literature Context (Non-Exhaustive)

Video inpainting literature commonly treats temporal consistency as a primary modeling constraint, using propagation mechanisms, spatiotemporal Transformers, and joint optimization across frames. Representative baselines include STTN [6], FuseFormer [7], E2FGVI [8], and ProPainter [9]. Diffusion-based video inpainting and editing has explored temporal consistency through feature reuse or attention control, for example Dreamix [10], FateZero [11], vid2vid-zero [12], Video-P2P [13], TokenFlow [14], Rerender-A-Video [15], and RAVE [16]. StableV2V is a recent example focusing on shape-consistent video-to-video editing and provides additional benchmark material [17].

VII. Recommendations for Strengthening the Work

1) Add baseline comparisons against at least one video-native inpainting method (for example STTN [6], FuseFormer [7], E2FGVI [8], ProPainter [9]) and at least one diffusion-based video editing consistency method (for example Dreamix [10], TokenFlow [14], Video-P2P [13]).
2) Define an explicit evaluation protocol: specify video sources, resolutions, clip lengths, mask generation procedure, prompts, sampling settings, random seeds, and compute hardware.
3) Report quantitative metrics alongside qualitative grids. Where applicable, include SSIM [23] and a perceptual metric such as LPIPS [22]. Add at least one temporal stability measure with a clear definition, for example the warping-consistency error in (7) evaluated over non-occluded regions.
4) Specify the occlusion-handling update rule for optical-flow warping, including how occluded regions are detected and how they are filled or blended across time.
5) Consider replacing purely per-frame mask generation with temporally coherent segmentation or propagation modules (for example SAM 2 [20] or DEVA [21]) and report the impact on temporal stability.

References

[1] S. Aghababyan, "AI-Powered Video Editing: Motion-Aware Object Replacement Using Generative Models," M.S. thesis, American University of Armenia, 2025.

[2] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," arXiv:2003.12039, 2020. [Online]. Available: https://arxiv.org/abs/2003.12039

[3] R. Rombach *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. CVPR*, 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[4] L. Zhang *et al.*, "Adding Conditional Control to Text-to-Image Diffusion Models," in *Proc. ICCV*, 2023 (ControlNet). [Online]. Available: https://arxiv.org/abs/2302.05543

[5] H. Ye *et al.*, "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models," arXiv:2308.06721, 2023. [Online]. Available: https://arxiv.org/abs/2308.06721

[6] Y. Zeng *et al.*, "Learning Joint Spatial-Temporal Transformations for Video Inpainting," in *Proc. ECCV*, 2020 (STTN). [Online]. Available: https://arxiv.org/abs/2007.10247

[7] R. Liu *et al.*, "FuseFormer: Fusing Fine-Grained Information in Transformers for Video Inpainting," in *Proc. ICCV*, 2021. [Online]. Available: https://arxiv.org/abs/2109.02974

[8] Z. Li *et al.*, "Towards an End-to-End Framework for Flow-Guided Video Inpainting," in *Proc. CVPR*, 2022 (E2FGVI). [Online]. Available: https://arxiv.org/abs/2204.02663

[9] S. Zhou *et al.*, "ProPainter: Improving Propagation and Transformer for Video Inpainting," in *Proc. ICCV*, 2023. [Online]. Available: https://arxiv.org/abs/2309.03897

[10] E. Molad *et al.*, "Dreamix: Video Diffusion Models are General Video Editors," arXiv:2302.01329, 2023. [Online]. Available: https://arxiv.org/abs/2302.01329

[11] C. Qi *et al.*, "FateZero: Fusing Attentions for Zero-shot Text-based Video Editing," in *Proc. ICCV*, 2023. [Online]. Available: https://arxiv.org/abs/2303.09535

[12] W. Wang *et al.*, "Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models (vid2vid-zero)," arXiv:2303.17599, 2023. [Online]. Available: https://arxiv.org/abs/2303.17599

[13] S. Liu *et al.*, "Video-P2P: Video Editing with Cross-attention Control," in *Proc. CVPR*, 2024. [Online]. Available: https://arxiv.org/abs/2303.04761

[14] M. Geyer *et al.*, "TokenFlow: Consistent Diffusion Features for Consistent Video Editing," in *Proc. ICLR*, 2024. [Online]. Available: https://arxiv.org/abs/2307.10373

[15] S. Yang *et al.*, "Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation," in *Proc. SIGGRAPH Asia*, 2023. [Online]. Available: https://arxiv.org/abs/2306.07954

[16] O. Kara *et al.*, "RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models," arXiv:2312.04524, 2023. [Online]. Available: https://arxiv.org/abs/2312.04524

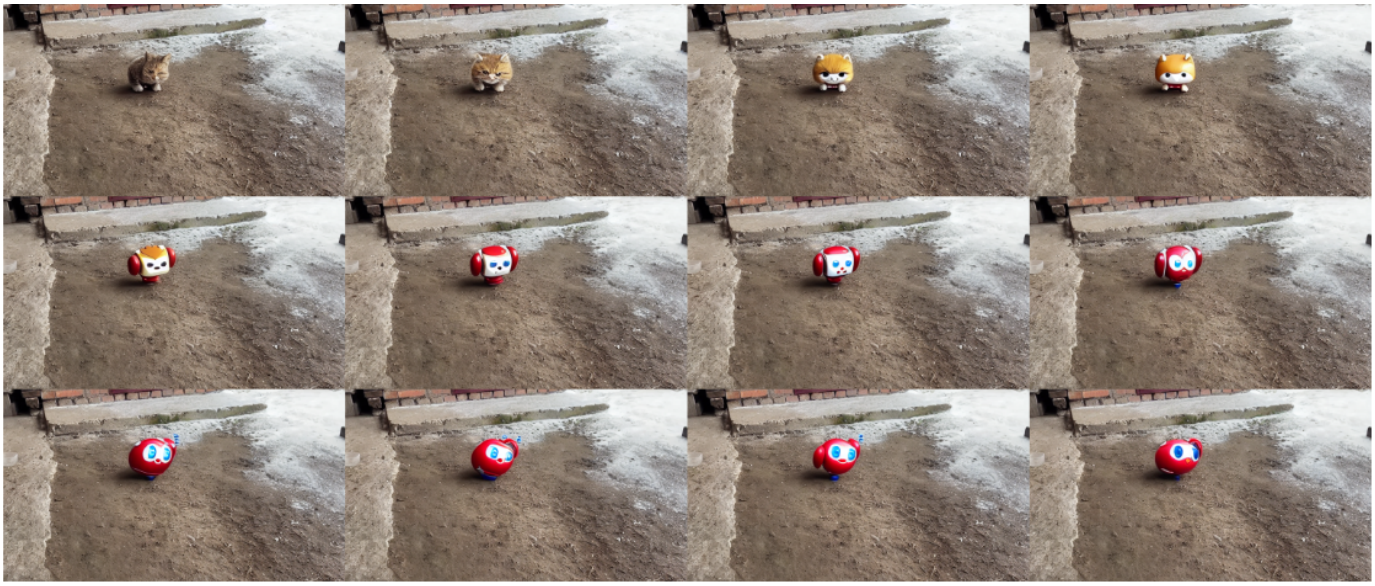

Fig. 5. Supplementary qualitative result from a reviewer-run notebook execution. Across the 12-frame sequence, the replacement remains near the intended scene region, but its appearance drifts strongly over time, changing color, shape, and apparent identity. The grid therefore supports the review's central claim that coarse motion transfer or location consistency alone does not guarantee temporal appearance consistency.

[17] C. Liu *et al.*, "StableV2V: Stablizing Shape Consistency in Video-to-Video Editing," arXiv:2411.11045, 2024. [Online]. Available: https://arxiv.org/abs/2411.11045

[18] LOVEU@CVPR'23, "Track 4: Text-Guided Video Editing (TGVE) Competition," 2023. [Online]. Available: https://sites.google.com/view/loveucvpr23/track4

[19] ShowLab, "LOVEU-TGVE 2023: Text-Guided Video Editing Competition (data, baselines, submission)," 2023. [Online]. Available: https://github.com/showlab/loveu-tgve-2023

[20] N. Ravi *et al.*, "SAM 2: Segment Anything in Images and Videos," arXiv:2408.00714, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[21] H. K. Cheng *et al.*, "Tracking Anything with Decoupled Video Segmentation (DEVA)," in *Proc. ICCV*, 2023. [Online]. Available: https://arxiv.org/abs/2309.03903

[22] R. Zhang *et al.*, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. CVPR*, 2018 (LPIPS). [Online]. Available: https://arxiv.org/abs/1801.03924

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] Y. Wu *et al.*, "VLA-AN: An Efficient and Onboard Vision-Language-Action Framework for Aerial Navigation in Complex Environments," arXiv:2512.15258, 2025. [Online]. Available: https://arxiv.org/abs/2512.15258 [Online]. Also available: https://ieeexplore.ieee.org/document/11297793

[25] Q. Chen, N. Gao, S. Huang, J. Low, T. Chen, J. Sun, and M. Schwager, "GRaD-Nav++: Vision-Language Model Enabled Visual Drone Navigation with Gaussian Radiance Fields and Differentiable Dynamics," arXiv:2506.14009, 2025. [Online]. Available: https://arxiv.org/abs/2506.14009