Monteinstitute : Pages:1–21

Original Research

Region-of-Interest Aware Diffusion Models for Controllable Video Editing

Ahsan Raza Siddiqui[1] and Muhammad Taha Qureshi[2]

[1]Department of Information Technology, Karakoram International University, University Road, Gilgit 15100, Gilgit-Baltistan, Pakistan.
[2]Department of Software Engineering, Mohammad Ali Jinnah University, Street 10, Phase II, Gulshan-e-Iqbal, Karachi 75300, Pakistan.

Abstract
Diffusion-based generative models have recently become a prominent approach for controllable image and video synthesis, enabling a range of applications in creative production, content retargeting, and post-processing workflows. These models typically operate over high-dimensional spatiotemporal tensors and rely on iterative denoising processes guided by conditioning signals such as text or exemplars. However, most existing approaches treat the video volume in a spatially uniform manner, which limits their ability to perform localized, semantically meaningful edits that preserve contextual consistency outside a user-specified region. This is a significant limitation in practical video editing scenarios, where users often require precise modifications within a region of interest while maintaining global coherence. This paper investigates region-of-interest aware diffusion models for controllable video editing, in which user-specified spatial or spatiotemporal regions guide the evolution of the denoising process. The proposed formulation treats regions of interest as first-class conditioning objects that influence sampling dynamics, attention patterns, and loss weighting. A tensorial representation of region masks is integrated into the diffusion process to jointly regulate spatial focus, temporal consistency, and identity preservation outside the edited areas. The study explores both training-time and sampling-time mechanisms for region control, including weighted reconstruction objectives and region-aware score fields. Experimental analyses on diverse editing tasks, including object replacement, attribute modification, and localized stylization, indicate that region-of-interest aware diffusion provides controllable behavior while maintaining temporal stability and content preservation in non-edited regions.

1. Introduction

Diffusion models have emerged as a flexible probabilistic framework for generative modeling of high-dimensional signals, particularly for visual modalities such as images and videos [1]. These models construct a Markovian or continuous-time noising process that gradually corrupts data into a simple reference distribution and learn a reverse process that incrementally denoises the corrupted samples. The reverse dynamics are parameterized by deep neural networks, typically U-shaped architectures equipped with multi-scale convolutions and attention mechanisms. When conditioned on auxiliary signals such as natural language descriptions, categorical labels, or reference frames, diffusion models support a variety of conditional synthesis and editing tasks. In the video setting, this flexibility enables controlled generation over spatiotemporal volumes, making such models attractive for editing tasks that must respect both spatial detail and temporal continuity.

Practical video editing workflows, however, rarely require global modification of every pixel. Instead, artists and downstream applications often operate on localized regions that encode objects, parts, or semantic entities, while requiring that the remainder of the scene remains unchanged both visually and temporally. Existing diffusion-based video editing methods typically introduce editing constraints at the global level, for example through prompt-based guidance, keyframe supervision, or spatially uniform cross-attention modulation. Although these strategies can influence the overall appearance or motion, they often lack precise control over where in the video specific changes occur, and they may inadvertently introduce unintended modifications in regions that are meant to remain fixed. This can be problematic for tasks such as selective replacement of objects, localized stylization of foreground actors, or editing of a particular temporal segment without altering the broader context.

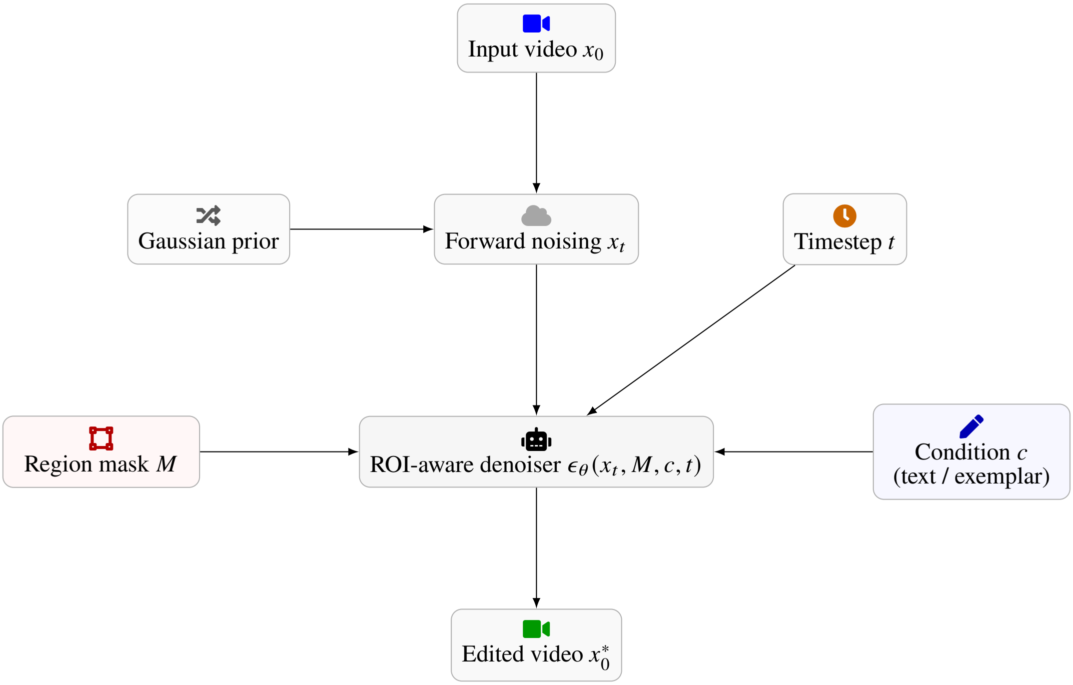

Figure 1: Overview of the region-of-interest aware video diffusion pipeline. The original video is diffused to a noisy latent, which is then denoised by a mask- and condition-aware network to produce an edited video that respects both the region constraints and the global temporal structure.

Symbol	Description	Shape / domain
X	Video tensor representation	$\mathbb{R}^{T\times H\times W\times C}$
x_0	Vectorized clean video sample	$\mathbb{R}^d$, $d = THWC$
$\{x_t\}_{t=0}^K$	Noisy samples along diffusion trajectory	$\mathbb{R}^d$ for each t
M	Region-of-interest mask (hard or soft)	$\{0, 1\}^{T\times H\times W}$ or $[0, 1]^{T\times H\times W}$
m	Vectorized mask	$\mathbb{R}^{d'}$, $d' = THW$
$P_{\text{roi}}, P_{\text{bg}}$	ROI / background projection operators	$x \mapsto m \odot x$, $x \mapsto (1 - m) \odot x$

Table 1: Key notation for video tensors, masks, and projection operators used in region-aware video diffusion models.

Region-of-interest representations provide an intuitive abstraction for specifying where edits should take effect. A region of interest can be defined as a spatial mask per frame, a spatiotemporal volume, or a per-voxel importance weighting field that indicates the strength of desired modifications. Integrating such region-aware signals into diffusion-based video editing requires rethinking how the noise prediction

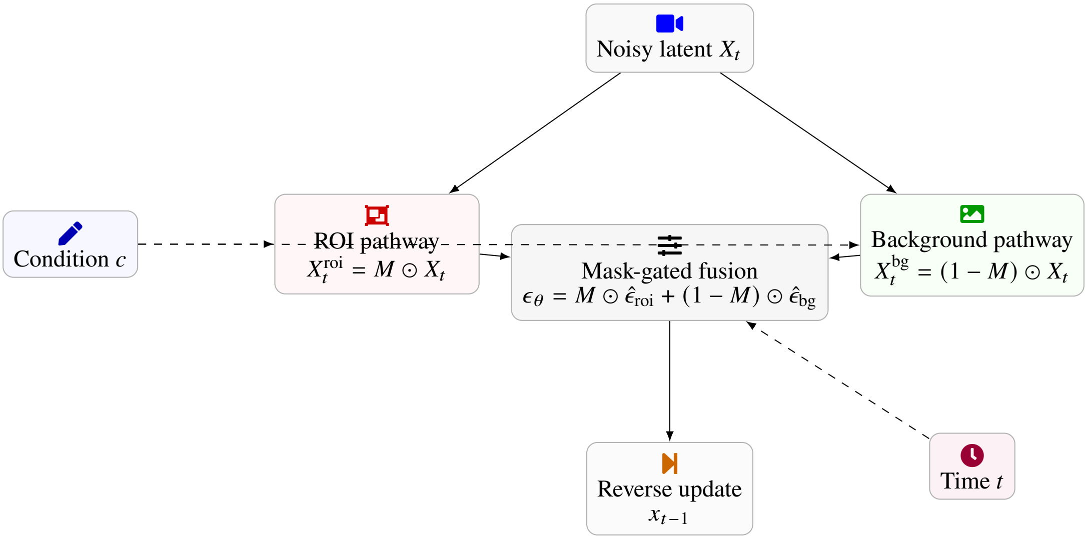
Noisy latent X_t
Condition c
ROI pathway
$X_t^{\text{roi}} = M \odot X_t$
Mask-gated fusion
$\epsilon_\theta = M \odot \hat{\epsilon}_{\text{roi}} + (1 - M) \odot \hat{\epsilon}_{\text{bg}}$
Background pathway
$X_t^{\text{bg}} = (1 - M) \odot X_t$
Reverse update
x_{t-1}
Time t

Figure 2: Mask-gated decomposition of the denoiser into region-of-interest and background components. Noisy latents are split by the mask, processed along two pathways, and recombined via a spatially varying mixture that yields different noise predictions for edited and preserved regions.

Quantity	Role in the model	Typical instantiation
$\{x_t\}$	Forward / reverse diffusion states	$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\ \epsilon$
ϵ_θ	Noise or score predictor	U-shaped spatiotemporal network with attention
c	Conditioning signal for controllable generation	Text prompt, labels, reference frames, style codes
$\{\beta_t\}, \{\alpha_t\}$	Variance / noise-retention schedule	Discrete schedule defining noising and denoising strength
$A(x_t)$	Spatiotemporal self-attention operator	Attention over flattened (t, h, w) tokens
$f(t), g(t)$	Drift and diffusion in SDE view	Scalar functions defining continuous-time dynamics

Table 2: Core diffusion and conditioning variables for video, connecting discrete-time updates and continuous-time stochastic formulations.

network, the sampling procedure, and the training objectives interact with spatially structured constraints. A direct application of image-based inpainting strategies to video may fail to account for temporal coherence, while naive masking of latents during denoising can lead to artifacts at region boundaries and discontinuities across frames. Therefore, a more principled formulation that jointly models the spatiotemporal structure and the region-aware conditioning is needed.

This paper studies diffusion models for controllable video editing that explicitly incorporate region-of-interest signals in both the model architecture and the probabilistic formulation. The central idea is to treat region masks as additional conditioning variables that modulate the denoising dynamics and the underlying score field in a spatially inhomogeneous manner. Rather than applying uniform noise removal across the entire video tensor, the model learns to adjust its reconstruction strength and editing behavior as a function of both the current noisy sample and the region specification. This leads to a framework where the video volume is decomposed into edited and preserved components, with different reconstruction objectives and sampling trajectories for each part.

The analysis begins by representing video data as high-order tensors in a Euclidean space and defining region masks as indicator or weighting tensors over the same index set. Building on this representation,

Mechanism	Where applied	Mask usage	Expected effect
Mask concatenation	Input to denoiser	M added as extra channels	Basic spatial localization of edits
Masked feature decomposition	Internal feature maps	Split into ROI / background branches	Different processing for edited vs. preserved areas
Mixture-of-experts noise head	Output of network	$\epsilon_\theta = M \odot \hat{\epsilon}_{\text{roi}} + (1 - M) \odot \hat{\epsilon}_{\text{bg}}$	Separate noise estimates blended by mask
Self-attention modulation	Spatiotemporal attention layers	Mask-dependent bias on attention logits	Stronger interactions inside ROI, controlled cross-talk
Cross-attention routing	Text or exemplar conditioning	Region-aligned relevance matrix R	Different spatial zones attend to different condition tokens
Hierarchical mask pyramids	Multi-scale U-Net levels	Downsampled $M^{(s)}$ at each scale	Coherent region reasoning from coarse to fine detail

Table 3: Region-of-interest conditioning mechanisms integrated into the denoising network architecture at different depths and modules.

Term	Definition	Impact on editing behaviour
s_{ij}	Baseline attention score $\frac{q_i^\top k_j}{\sqrt{d_k}}$	Standard spatiotemporal token interactions
$\gamma\, m_i m_j$	ROI–ROI bias	Encourages stronger coupling within the edited region
$\delta\, m_i(1 - m_j)$	ROI → background bias	Controls influence of edits on preserved areas
$\eta\, (1 - m_i)m_j$	Background → ROI bias	Limits leakage of background structure into ROI edits
a_{ij}	Attention weight after softmax over $\tilde{s}_{ij}$	Final region-aware aggregation of value vectors v_j

Table 4: Mask-aware self-attention terms that shape interactions within and across regions of interest during denoising.

Design element	Mathematical form	Intuitive effect
Spatial guidance field	$\epsilon_{\text{roi}} = \epsilon_\theta(x_t, \varnothing) + g \odot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing))$	Concentrates classifier-free guidance inside ROI
Diagonal preconditioner	$W = \text{diag}(w_i)$ with larger w_i in ROI	Stronger local update steps where edits are desired
Time-varying guidance	$g_t = \gamma_t M$ with γ_t increasing as $t \to 0$	Coarse global structure early, sharper edits late
Dynamic mask evolution	$M_{t-1} = \Psi(M_t, x_t, c)$	Allows ROI to dilate or erode in response to generated content
Smoothed boundary mask	$\tilde{M} = M + \alpha\Delta\tilde{M}$	Soft transition layer reducing artifacts at region edges

Table 5: Region-dependent guidance and preconditioning strategies that induce non-uniform reverse dynamics across the video volume.

the diffusion process is extended with region-aware operators that project the global score field onto subspaces associated with edited and non-edited regions [2]. The noise prediction network receives both the noisy video latent and the region tensor as inputs, and its architecture is adapted to exploit region structure through spatial gating and attention re-weighting. At training time, loss terms with different weights are assigned to region interior, region boundaries, and background to balance edit fidelity and identity preservation. At sampling time, the region signal influences the reverse dynamics via guidance mechanisms that selectively emphasize edits within the specified spatial support.

The proposed region-of-interest aware diffusion formulation is evaluated on several types of controllable video editing tasks. These tasks involve modifying attributes of objects, replacing or inserting

Loss term	Region focus	Objective	Weight symbol
$\mathcal{L}_{\text{roi}}$	Inside ROI	Masked denoising error on ϵ	λ_{roi}
$\mathcal{L}_{\text{bg}}$	Background	Masked denoising error outside ROI	λ_{bg}
$\mathcal{L}_{\text{edit}}$	Inside ROI	Reconstruction of edited target x'_0 (e.g., ℓ_1)	η_{edit}
$\mathcal{L}_{\text{preserve}}$	Background	Reconstruction of original x_0 (e.g., ℓ_1)	η_{preserve}
$\mathcal{L}_{\text{temp}}$	Temporal neighbours	Smoothness of frame-wise features or pixels	η_{temp}
$\mathcal{L}_{\text{reg}}$	Global	Regularization of parameters (e.g., weight decay)	η_{reg}
$\mathcal{L}_{\text{total}}$	Global	Weighted sum of all components	–

Table 6: Training objectives decomposed into region-aware denoising, editing, preservation, and temporal consistency terms.

Sampling scheme	Description	Pros and trade-offs
DDPM-style Euler updates	Single-step explicit update per timestep	Simple, stable, may require many steps for high quality
Second-order Runge–Kutta	Two evaluations per step with intermediate state	Better local accuracy, fewer steps at higher cost per step
Two-phase guidance schedule	Weak guidance at high noise, strong near $t = 0$	Balances global structure with precise local edits
Noising original video	Initialize with $x_T = \sqrt{\alpha_T} x_0 + \sqrt{1 - \alpha_T}\epsilon$	Anchors background to input while allowing ROI changes
Region-focused computation	Restrict expensive modules to ROI neighbourhood	Lower cost when regions are small, requires careful implementation

Table 7: Sampling and numerical schemes tailored to region-aware video editing in diffusion models.

Aspect	Metric	Region focus	What it probes
Background preservation	Pixel or perceptual distance to x_0	Outside ROI	Identity and context stability in non-edited areas
Edit alignment	Classifier or feature-based attribute score	Inside ROI	Strength and correctness of applied edit or style
Temporal coherence	Temporal gradient magnitude G_{temp}	ROI vs. background	Flicker, oversmoothing, and motion consistency
Boundary quality	Error in boundary band E_{bound}	Around ∂M	Artifacts and blending at ROI interfaces
Latent structure	Covariances $\Sigma_{\text{roi}}(t), \Sigma_{\text{bg}}(t)$	Feature space	Allocation of variation to edited vs. preserved regions
Mask robustness	Sensitivity S to small mask perturbations	Near ROI borders	Stability under imperfect segmentations or mask edits

Table 8: Evaluation axes and quantitative metrics for analyzing controllability, coherence, and robustness of region-aware video diffusion models.

new content, and applying localized style transformations. The experiments focus on analyzing controllability, preservation of identity outside the edited region, and temporal smoothness of both edited and non-edited areas. While the study does not exhaustively optimize architectures or hyperparameters, it examines how different choices of region embedding, loss weighting, and sampling schemes affect the trade-off between edit precision and global coherence. The results suggest that region-aware constraints offer a useful means for structuring the generative process in video diffusion models, particularly when coupled with spatiotemporal attention mechanisms that respect the geometry of the video tensor.

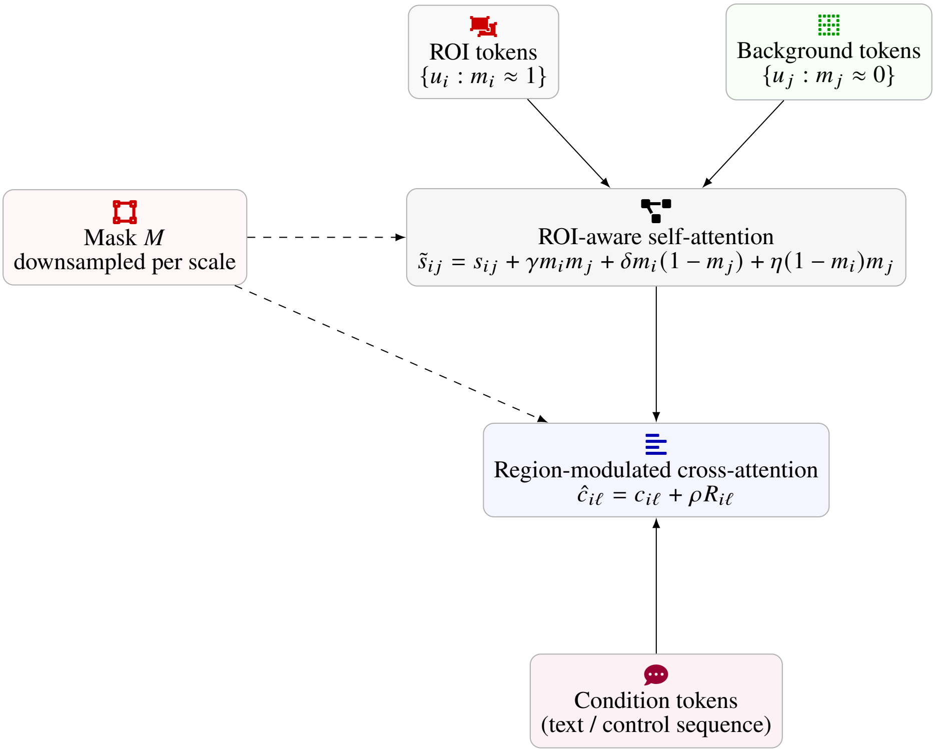
ROI tokens
$\{u_i : m_i \approx 1\}$
Background tokens
$\{u_j : m_j \approx 0\}$
Mask M
downsampled per scale
ROI-aware self-attention
$\tilde{s}_{ij} = s_{ij} + \gamma m_i m_j + \delta m_i(1 - m_j) + \eta(1 - m_i)m_j$
Region-modulated cross-attention
$\hat{c}_{i\ell} = c_{i\ell} + \rho R_{i\ell}$
Condition tokens
(text / control sequence)

Figure 3: Region-aware attention mechanisms. Self-attention logits are modulated by the spatial mask to emphasize interactions within the edited region and regulate cross-region influence, while cross-attention is biased by a region-to-token association matrix aligning spatial locations with specific conditioning tokens.

2. Background on Diffusion Models and Video Representations

In diffusion-based generative modeling, a data sample is treated as a random vector in a high-dimensional Euclidean space and is progressively perturbed by a forward noising process. For video, consider a tensor representation

$$X \in \mathbb{R}^{T\times H\times W\times C},$$

where T denotes the number of frames, H and W are spatial dimensions, and C is the number of channels. For convenience, this tensor can be reshaped into a vector $x_0 \in \mathbb{R}^d$ with

$$d = THWC.$$

The forward diffusion process defines a sequence of random variables $\{x_t\}_{t=0}^{K}$ that gradually add Gaussian noise to the original data. In a discrete-time formulation with variance schedule $\{\beta_t\}$, the forward transitions are defined as

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1 - \alpha_t}\, \epsilon, \tag{2.1}$$

where α_t is the cumulative product of noise retention factors and ϵ is a standard Gaussian random vector in $\mathbb{R}^d$. This closed-form expression for x_t in terms of x_0 allows efficient sampling of noisy data at arbitrary diffusion steps without explicitly simulating each intermediate transition. The reverse process seeks to approximate the conditional distribution $p(x_{t-1} \mid x_t)$ using a learned neural network that predicts either the original data, the noise, or the local score.

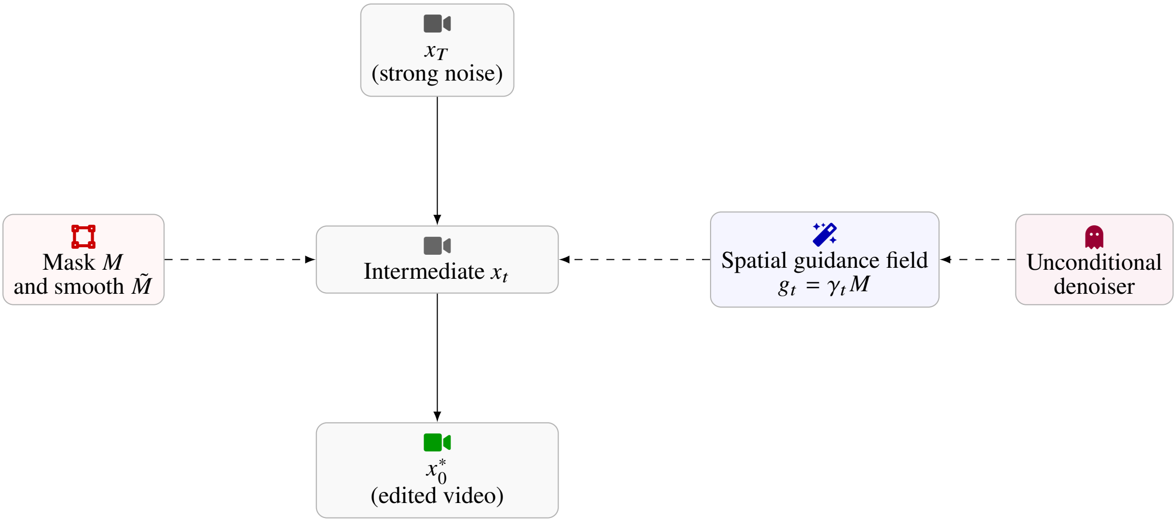

Figure 4: Spatiotemporal denoising with spatially varying guidance. A guidance field derived from the region mask scales the difference between conditional and unconditional denoisers, concentrating edit strength in the region while retaining unconditional reconstruction in the background along the reverse diffusion trajectory.

For controllable generation, the reverse dynamics are conditioned on auxiliary information c, which may be text, semantic labels, reference images, or other control signals. A common parameterization introduces a neural network ϵ_θ that predicts the noise component given x_t, the timestep t, and the condition c. The reverse update can then be expressed as

$$x_{t-1} = a_t x_t + b_t\, \epsilon_\theta(x_t, t, c) + \sigma_t z_t, \tag{2.2}$$

where a_t, b_t, and σ_t are scalar coefficients determined by the noise schedule and z_t is a fresh Gaussian noise term. The network ϵ_θ is usually instantiated as a U-shaped architecture with multi-scale convolutions that process spatial dimensions and temporal convolutions or attention layers that propagate information across frames.

When extending diffusion models to video, the tensor representation introduces considerable computational challenges. The dimensionality d grows linearly with the temporal dimension, and naive application of isotropic convolutions or dense attention across the entire spatiotemporal volume can be prohibitively expensive. To manage this complexity, many video diffusion architectures adopt factorized structures, for example applying two-dimensional convolutions per frame and augmenting them with lightweight temporal modules or low-rank attention in time [3]. In the context of region-of-interest aware modeling, this factorization can be leveraged to inject spatially localized information, since operations in the spatial dimensions are naturally aligned with the notion of regions.

Region representations in video can be formalized as tensors

$$M \in \{0, 1\}^{T\times H\times W},$$

where $M_{t,h,w} = 1$ indicates that the spatial position (h, w) in frame t belongs to the region of interest, and $M_{t,h,w} = 0$ otherwise. More general soft masks can be described by values in $[0, 1]$ that describe the relative importance of editing each spatiotemporal location. For mathematical analysis, it is convenient to vectorize the mask into $m \in \mathbb{R}^{d'}$, where $d' = THW$, and interpret it as a diagonal weighting operator on the ambient space of video pixels. Define the linear operator P_{roi} acting on a vectorized video x by

$$P_{\mathrm{roi}} x = m \odot x, \tag{2.3}$$

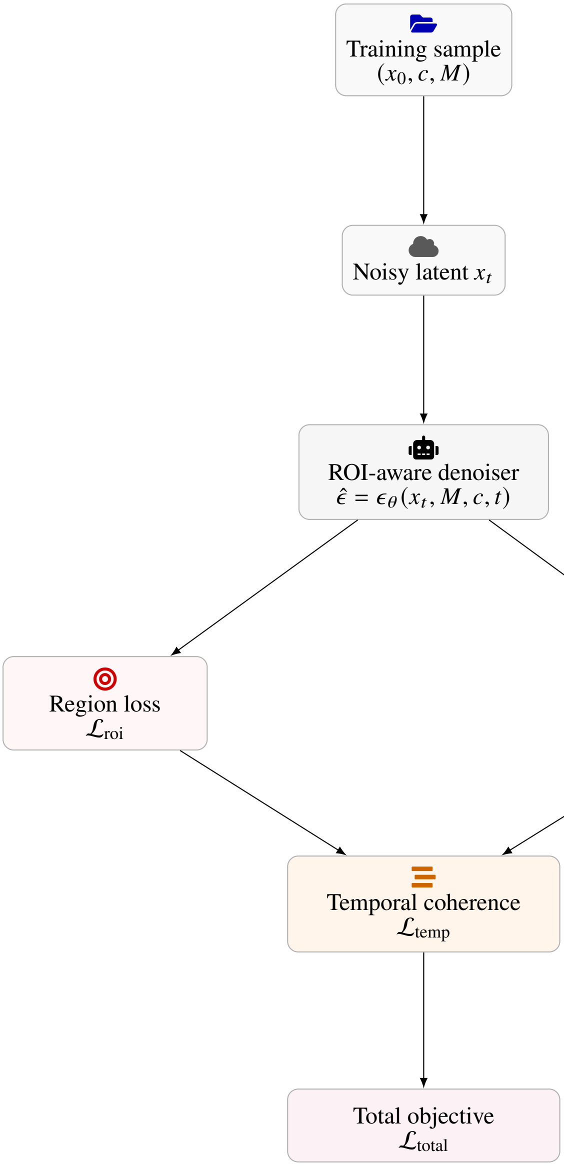
Training sample
(x_0, c, M)
Noisy latent x_t
ROI-aware denoiser
$\hat{\epsilon} = \epsilon_\theta(x_t, M, c, t)$
Region loss
$\mathcal{L}_{\text{roi}}$
Background loss
$\mathcal{L}_{\text{bg}}$
Temporal coherence
$\mathcal{L}_{\text{temp}}$
Total objective
$\mathcal{L}_{\text{total}}$

Figure 5: Structure of the training objective. The denoiser is supervised by separate reconstruction losses in the region and background, which are combined with temporal coherence regularization to form a total loss that balances edit fidelity, identity preservation, and dynamics.

where $\odot$ denotes componentwise multiplication. The complementary operator P_{bg} corresponding to the background or non-edited region is then

$$P_{\text{bg}}x = (1 - m) \odot x. \tag{2.4}$$

These operators decompose the video into edited and preserved components within the same ambient space.

In probabilistic terms, one can view the joint distribution of the full video and the mask as a random pair (x_0, m). For most editing scenarios, the mask is provided by a user or an external system and is treated as a fixed conditioning variable. The goal of region-of-interest aware video editing is to generate

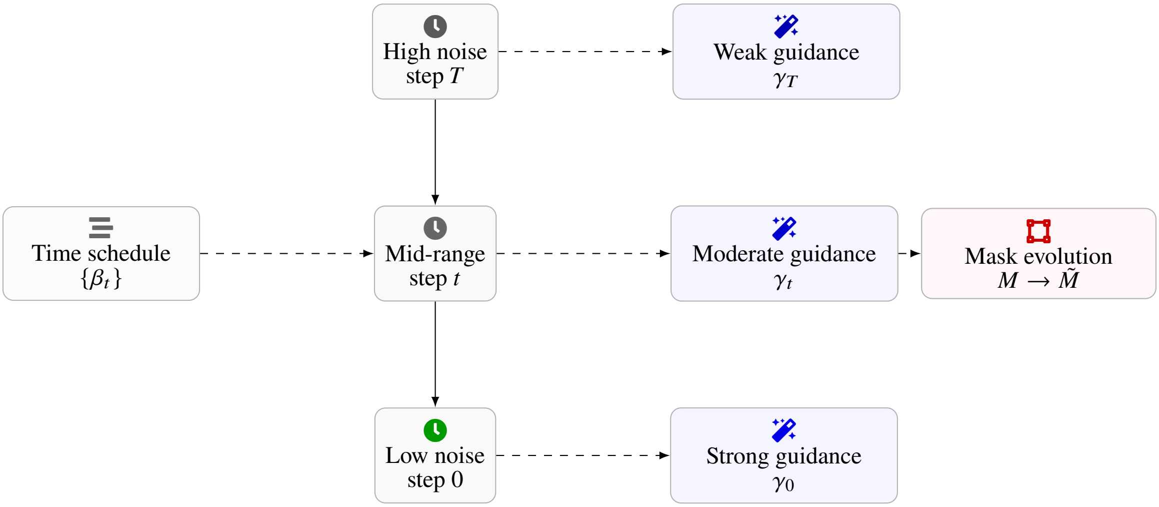
High noise step T
Weak guidance γ_T
Time schedule $\{\beta_t\}$
Mid-range step t
Moderate guidance γ_t
Mask evolution $M \to \tilde{M}$
Low noise step 0
Strong guidance γ_0

Figure 6: Time-dependent control of the reverse process. Guidance strength is increased toward low-noise steps, while mask smoothing or evolution can be applied in mid-range steps, allowing coarse global structure to form before enforcing strong localized edits near the end of sampling.

a modified video x_0^* that satisfies two constraints. First, the edited region should conform to the desired control specified by a condition c, which may encode a target object, style, or motion [4]. Second, the non-edited region should remain close to the original content to preserve identity and context. This can be expressed informally as

$$P_{\text{roi}} x_0^* \approx \tilde{x}_{\text{roi}}(c), \tag{2.5}$$

$$P_{\text{bg}} x_0^* \approx P_{\text{bg}} x_0, \tag{2.6}$$

where $\tilde{x}_{\text{roi}}(c)$ is an implicit target determined by the condition. In practice, $\tilde{x}_{\text{roi}}(c)$ may not be available as an explicit tensor and is instead implemented through guidance signals or learned representations. The diffusion model must reconcile these goals while respecting the temporal structure induced by the sequence of frames.

A common way to encode temporal coherence in video diffusion models is through spatiotemporal attention mechanisms. Let the vectorized representation of the video at step t be x_t. Self-attention defines an interaction kernel

$$A(x_t) = \operatorname{softmax}\left(\frac{Q(x_t)K(x_t)^\top}{\sqrt{d_k}}\right)V(x_t), \tag{2.7}$$

where Q, K, and V are learned linear projections and d_k is the key dimension [5]. When applied to flattened spatiotemporal tokens, attention can propagate information across all pixels and frames. For region-of-interest aware modeling, this mechanism offers an opportunity to bias interactions towards regions, by modifying either the attention logits or the value aggregation according to the mask. Introducing the mask as an additional factor can transform the global attention into a spatially modulated operation that emphasizes or de-emphasizes certain regions during denoising.

From the perspective of stochastic calculus, discrete-time diffusion models approximate the time-discretized solution of a reverse-time stochastic differential equation. In the continuous-time framework, the forward process is defined by

$$\mathrm{d}x = f(t)x\,\mathrm{d}t + g(t)\,\mathrm{d}w_t, \tag{2.8}$$

where w_t is a Wiener process and f and g are scalar functions specifying the drift and diffusion coefficients. The reverse-time dynamics have the form

$$\mathrm{d}x = \left(f(t)x - g(t)^2 \nabla_x \log p_t(x)\right) \mathrm{d}t + g(t)\, \mathrm{d}\bar{w}_t, \tag{2.9}$$

where $p_t(x)$ is the marginal density at time t and $\bar{w}_t$ is a Wiener process in reverse time. In this view, the neural network approximates the score field $\nabla_x \log p_t(x \mid c)$. Region-of-interest aware modeling can be interpreted as an approximation of a spatially inhomogeneous score field in which different regions of the domain have different prior structures or conditional behaviors. By designing the network and objective to respect the decomposition induced by the mask, one can bias the score estimator to behave differently in edited and non-edited regions, thus implementing controllable video editing via the generative process itself [6] [7].

3. Region-of-Interest Aware Conditioning Mechanisms

Incorporating region-of-interest signals into diffusion-based video editing requires a modeling strategy that can handle spatial localization, temporal consistency, and compatibility with the existing denoising network. A basic approach is to concatenate the mask M with the video latent as an additional channel and allow the network to infer how to use this information. This strategy treats the mask as an image-like input and permits convolutional layers to propagate mask information through spatial neighborhoods. However, such a simple concatenation may not be sufficient to achieve fine controllability, especially when the network must differentiate sharply between edited and preserved regions while maintaining temporal coherence and visual consistency at region boundaries.

A more structured approach can be formulated by treating the mask as an operator that acts on both the input and the intermediate feature maps. Consider the video latent at time step t, represented as a tensor

$$X_t \in \mathbb{R}^{T\times H\times W\times C}.$$

Let M be a soft mask with values in $[0, 1]$. Define

$$X_t^{\mathrm{roi}} = M \odot X_t, \tag{3.1}$$

$$X_t^{\mathrm{bg}} = (1 - M) \odot X_t\,[8]. \tag{3.2}$$

These tensors isolate contributions from the region of interest and the background. The diffusion network can be partitioned conceptually into two pathways, one focusing on X_t^{roi} and the other focusing on X_t^{bg}, with shared parameters or partially separate parameters. For example, an encoder can process the concatenated pair

$$\Phi_t = \phi\left(X_t^{\mathrm{roi}}, X_t^{\mathrm{bg}}, c, t\right), \tag{3.3}$$

where ϕ denotes a multi-stage neural transformation that incorporates temporal modules and attention layers, and the condition c is injected through adaptive normalization or cross-attention. The region-aware representation Φ_t is then decoded to produce a noise estimate $\hat{\epsilon}_{\mathrm{roi}}$ and $\hat{\epsilon}_{\mathrm{bg}}$ for region and background respectively. The final noise prediction can be assembled as

$$\epsilon_\theta(X_t, M, c, t) = M \odot \hat{\epsilon}_{\mathrm{roi}} + (1 - M) \odot \hat{\epsilon}_{\mathrm{bg}}. \tag{3.4}$$

This decomposition is a simple example of a mixture-of-experts structure, where the mask acts as a gating signal.

Region-of-interest information can also be integrated into attention mechanisms. Consider a self-attention layer operating on a sequence of tokens obtained by flattening the spatiotemporal grid of the video. Each token u_i corresponds to a particular (t, h, w) index and has an associated mask value m_i [9]. In dot-product attention, the unnormalized compatibility between tokens i and j is given by

$$s_{ij} = \frac{q_i^\top k_j}{\sqrt{d_k}}, \tag{3.5}$$

where q_i and k_j are query and key vectors. To encourage stronger interactions within the region of interest and controlled interactions between region and background, one can modify the scores as

$$\tilde{s}_{ij} = s_{ij} + \gamma\, m_i m_j + \delta\, m_i(1 - m_j) + \eta\,(1 - m_i)m_j, \tag{3.6}$$

where γ, δ, and η are learnable or hand-tuned scalars that regulate attention within and across regions. When γ is positive and larger in magnitude than δ and η, the network is encouraged to correlate positions inside the region more strongly, thereby enhancing the coherence of edits across spatial locations within the region. If δ and η are negative, the attention between region and background can be softly suppressed, reducing unintended influence of edits on non-target areas. These modulated scores are passed through a softmax to produce attention weights, yielding

$$a_{ij} = \frac{\exp(\tilde{s}_{ij})}{\sum_k \exp(\tilde{s}_{ik})}. \tag{3.7}$$

The resulting weighted sum over value vectors v_j generates updated token representations that respond differently to region and background, while preserving differentiability with respect to both network parameters and, in the case of soft masks, the mask values [10].

Beyond self-attention, cross-attention layers used for conditioning on text or other signals offer another locus for integrating region-of-interest information. Let y_ℓ denote a token corresponding to the conditioning sequence, such as a text embedding. Cross-attention typically computes

$$c_{i\ell} = \frac{q_i^\top k_\ell}{\sqrt{d_k}}, \tag{3.8}$$

where q_i depends on the video token and k_ℓ depends on the condition. If a prompt contains multiple semantic phrases, one can align specific phrases with specific regions and modulate cross-attention accordingly. This can be achieved by introducing a matrix $R \in \mathbb{R}^{N\times L}$, where N is the number of video tokens and L is the number of condition tokens, and letting $R_{i\ell}$ encode the relevance of phrase ℓ to the spatial position of token i. The attention logits can then be adjusted as

$$\hat{c}_{i\ell} = c_{i\ell} + \rho\, R_{i\ell}, \tag{3.9}$$

with a scaling parameter ρ. When regions are specified for different semantic entities, the entries of R can be derived from the masks associated with each entity, thus providing fine-grained spatial control during conditioning. The probability distribution over condition tokens for each video token becomes

$$\alpha_{i\ell} = \frac{\exp(\hat{c}_{i\ell})}{\sum_k \exp(\hat{c}_{ik})}, \tag{3.10}$$

and the context vector for token i is the weighted sum of condition value vectors. This mechanism allows different parts of the video to attend to different components of the conditioning sequence in a manner aligned with region annotations.

On the numerical side, region-of-interest aware conditioning can be interpreted as inducing a non-isotropic noise removal process [11]. In the standard diffusion setting, the denoiser attempts to remove noise uniformly across all coordinates, reflecting the assumption that all pixels have identical prior distributions. Introducing region-specific behavior effectively imposes a coordinate-wise preconditioner on the denoising process. Let $W \in \mathbb{R}^{d\times d}$ be a diagonal matrix with entries w_i that encode the relative strength of reconstruction at each pixel, where w_i may depend on the mask. The reverse update can then be written as

$$x_{t-1} = x_t + \Delta t\, W\,(\mu_\theta(x_t, M, c, t) - x_t) + \sigma_t z_t, \tag{3.11}$$

where μ_θ is a predicted mean and Δt is a step size. If the weights w_i are larger for pixels in the region of interest, the reverse dynamics will adjust those coordinates more strongly at each step, effectively amplifying the influence of editing constraints within the region compared to the background. This preconditioning perspective provides a link between mask-aware conditioning and the dynamical system underlying the diffusion model [12].

To ensure that region-aware conditioning remains compatible with the global coherence of the video, it is necessary to consider how mask information propagates through the hierarchy of scales in the network. In U-shaped architectures for video, feature maps at multiple resolutions represent progressively coarser views of the spatiotemporal volume. The mask M can be downsampled using average pooling or learned projections to obtain masks at each resolution level, enabling the network to reason about regions both at fine detail and coarse structure. If the mask is treated as a continuous function over spatiotemporal coordinates, one can view this downsampling as approximating the convolution

$$M^{(s)}(t, h, w) = \int K_s(u, v, \tau) M(t+\tau, h+u, w+v)\mathrm{d}u\, \mathrm{d}v\, \mathrm{d}\tau, \tag{3.12}$$

where K_s is a scale-dependent kernel. This continuous interpretation suggests that mask signals can be smoothed and aggregated across neighborhoods to capture region shape and context at various scales, which is beneficial for editing operations that must respect the geometry of objects and their surroundings.

4. Spatiotemporal Denoising with Controllable Editing Constraints

Region-of-interest aware video editing can be viewed as solving a constrained generative problem, where the constraints reflect both the desired edits in the region and the requirement that the background remain close to the original video. In probabilistic terms, given an input video x_0, a mask M, and a condition c, the goal is to sample from a conditional distribution [13] [14]

$$p_\theta(x_0^* \mid x_0, M, c), \tag{4.1}$$

such that x_0^* satisfies editing constraints within the region and preservation constraints outside it. One way to conceptualize this is to treat the edited video as the solution to an optimization problem over trajectories of the reverse diffusion process. Let $\{x_t\}_{t=0}^{K}$ denote a reverse trajectory with x_K sampled from a standard Gaussian prior. The objective can be written informally as minimizing an energy functional

$$\mathcal{E}(x_{0:K}) = \sum_{t=1}^{K} \ell_t(x_t, M, c; \theta) + \lambda_{\mathrm{bg}} \left\| P_{\mathrm{bg}} x_0 - P_{\mathrm{bg}} x_0^* \right\|_2^2, \tag{4.2}$$

subject to the constraint that $\{x_t\}$ follows the reverse transition dynamics induced by the diffusion model. Here, ℓ_t encapsulates the mismatch between the reverse transition and the learned denoiser at each step, while the second term enforces background preservation.

In practice, direct trajectory optimization is not performed. Instead, one modifies the reverse updates to incorporate guidance terms that reflect the desired constraints in a stepwise manner. A common form of classifier-free guidance adjusts the predicted noise as

$$\epsilon_{\mathrm{guided}} = \epsilon_\theta(x_t, t, \varnothing) + \gamma[15]\left(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \varnothing)\right), \tag{4.3}$$

where $\varnothing$ denotes the absence of conditioning and γ is a guidance scale. In the region-of-interest aware setting, this guidance can be made spatially varying. Define a guidance field $g \in \mathbb{R}^{d'}$ derived from the mask, taking values in $[0, 1]$. The guided noise prediction becomes

$$\epsilon_{\mathrm{roi}}(x_t, M, c, t) = \epsilon_\theta(x_t, t, \varnothing) + g \odot (\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \varnothing)) \,. \tag{4.4}$$

By setting g to be large in the region and small in the background, the edit-specific guidance is concentrated where desired, while the background is reconstructed more according to the unconditional model that approximates the original content distribution. This approach yields a spatially non-uniform guidance scheme that aligns with user-specified regions [16].

Temporal consistency is a key concern in video editing, particularly at the boundaries of the region where edited objects interact with unedited background. To reason about temporal coherence, consider the video as a function $x_0(t, h, w, c)$ defined on a discrete time axis and continuous spatial coordinates. Temporal differences across frames can be approximated by finite differences. For frame index k, the temporal derivative at position (h, w) can be approximated as

$$D_t x_0(k, h, w) = x_0(k + 1, h, w) - x_0(k, h, w). \tag{4.5}$$

For a consistent video, these differences should vary smoothly across time, especially along trajectories corresponding to the same physical point in the scene. In the presence of an edited region, one aims to maintain smoothness not only within the region but also across the interface between region and background. To enforce this, the diffusion model can incorporate a penalization of temporal gradients weighted by the mask [17]. Define a temporal smoothness functional

$$\mathcal{S}(x_0^*, M) = \sum_k \sum_{h,w} \omega_{k,h,w} \left\| D_t x_0^*(k, h, w) \right\|_2^2 , \tag{4.6}$$

where $\omega_{k,h,w}$ is a weight that depends on both the mask value and perhaps its spatial derivatives, increasing near region boundaries. During training, a surrogate version of this functional can be applied at intermediate diffusion steps by approximating temporal differences in the noisy latents and backpropagating gradients through the denoising network.

From the standpoint of multivariate calculus, the diffusion model defines a mapping

$$F_\theta : \mathbb{R}^d \times \mathbb{R}^{d'} \times C \times \{1, \ldots, K\} \to \mathbb{R}^d, \tag{4.7}$$

where C denotes the space of conditions, and

$$F_\theta(x_t, m, c, t) = \epsilon_\theta(x_t, m, c, t). \tag{4.8}$$

The associated reverse update can be seen as an explicit method for integrating an ordinary differential equation with additive noise. For small step sizes, this resembles a forward Euler discretization of a drift field

$$b(x_t, m, c, t) = -\kappa_t F_\theta(x_t, m, c, t), \tag{4.9}$$

with scaling κ_t depending on the noise schedule. The presence of the mask introduces an additional dependence of the drift field on spatial coordinates, effectively making the vector field non-homogeneous across the domain [18]. The Jacobian of this field with respect to x_t,

$$J_t = \nabla_{x_t} b(x_t, m, c, t), \tag{4.10}$$

determines the local stability and contraction properties of the reverse dynamics. If large eigenvalues of J_t are concentrated in the subspace associated with the region of interest, while the background subspace has smaller eigenvalues, the dynamics will adjust more strongly within the region. This observation connects the choice of training objectives and mask weighting with the numerical behavior of the sampling algorithm.

The interface between edited and preserved regions can be modeled using tools from discrete geometry. Let ∂M denote a discrete approximation of the boundary of the mask, defined as the set of voxels whose mask value differs from at least one neighbor. To avoid artifacts at ∂M, it is useful to relax the binary mask into a continuous transition layer. One can define a smoothed mask $\tilde{M}$ as the solution of a discrete diffusion equation on the spatial grid. For a frame k, consider the Laplacian operator

$$(\Delta\tilde{M})_{h,w} = \sum_{(u,v)\in\mathcal{N}(h,w)} \left(\tilde{M}_{u,v} - \tilde{M}_{h,w}\right), [19] \tag{4.11}$$

where $\mathcal{N}(h,w)$ is a neighborhood around (h,w). Solving

$$\tilde{M} = M + \alpha\Delta\tilde{M}, \tag{4.12}$$

for a small α yields a mask with softened boundaries. This smoothed mask can regulate the transition of editing strength from fully edited to fully preserved regions and can be used both in the network conditioning and in the loss weighting, reducing sharp discontinuities that might otherwise cause ringing or ghosting artifacts in the resulting video.

Another aspect of controllable editing is the ability to adjust the extent of the region over time or as a function of diffusion step. Users may specify an initial mask that is then dilated or eroded adaptively during sampling. Mathematically, this corresponds to evolving the mask under morphological operations that depend on the current denoising state. For instance, let M_t denote the mask at diffusion step t. One may define an update

$$M_{t-1} = \Psi(M_t, x_t, c), \tag{4.13}$$

where Ψ is a function that performs dilation in areas where the network predicts strong consistency between edited content and background or erodes regions where the edit conflicts with structural constraints [20]. While such dynamic mask evolution introduces additional complexity, it provides a mechanism for refining the region-of-interest during denoising in response to the generative process itself.

5. Numerical Optimization, Training Objectives, and Sampling Algorithms

Training a region-of-interest aware diffusion model for video editing requires designing an objective function that balances reconstruction fidelity, edit specificity, temporal coherence, and numerical stability. Let $\mathcal{D}$ denote a dataset of triplets (x_0, c, M) representing original videos, associated conditions, and masks. The conditions can encapsulate text prompts, reference images, or other control modalities. During training, one samples a timestep t, draws Gaussian noise ϵ, and constructs the noisy latent

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon. \tag{5.1}$$

The diffusion network receives x_t, the mask M, the condition c, and the timestep t, and outputs a noise prediction $\hat{\epsilon}$. The basic denoising objective minimizes a weighted mean squared error between ϵ and $\hat{\epsilon}$. In the region-aware setting, the error is decomposed according to the mask. Define

$$\mathcal{L}_{\text{roi}} = \mathbb{E}\left[\|M \odot (\epsilon[21] - \hat{\epsilon})\|_2^2\right], \tag{5.2}$$

$$\mathcal{L}_{\text{bg}} = \mathbb{E}\left[\|(1-M) \odot (\epsilon - \hat{\epsilon})\|_2^2\,[22]\right]. \tag{5.3}$$

These two losses can be combined with distinct weights

$$\mathcal{L}_{\text{denoise}} = \lambda_{\text{roi}}\mathcal{L}_{\text{roi}} + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}}, \tag{5.4}$$

where λ_{roi} and λ_{bg} regulate the trade-off between editing strength and background preservation. When training for editing, it is often desirable to place higher emphasis on accurate noise prediction inside the region than outside, leading to $\lambda_{\text{roi}} > \lambda_{\text{bg}}$.

In many editing scenarios, there is an explicit or implicit target video x'_0 that reflects the desired edited content. This can arise, for example, when training from pairs of before-and-after videos or when simulating edits via self-supervision. In such cases, one can introduce an additional reconstruction loss defined on the denoised sample after a full reverse trajectory. Let $\tilde{x}_0$ denote the reconstruction obtained by running the reverse process from a noisy sample. A supervised editing loss can be expressed as

$$\mathcal{L}_{\text{edit}} = \mathbb{E}\left[\left\|M \odot (\tilde{x}_0 - x'_0)\ [23]\right\|_1\right], \tag{5.5}$$

using an ℓ_1 norm to increase robustness to outliers. A background preservation loss can also be defined as

$$\mathcal{L}_{\text{preserve}} = \mathbb{E}\left[\|(1 - M) \odot (\tilde{x}_0 - x_0)\|_1\right] [24]. \tag{5.6}$$

These terms penalize deviations from the original content outside the region while encouraging alignment with the editing target inside the region. The total training loss combines these components with temporal smoothing and regularization terms,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \eta_{\text{edit}}\mathcal{L}_{\text{edit}} + \eta_{\text{preserve}}\mathcal{L}_{\text{preserve}} + \eta_{\text{temp}}\mathcal{L}_{\text{temp}} + \eta_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{5.7}$$

where $\mathcal{L}_{\text{temp}}$ enforces temporal coherence and $\mathcal{L}_{\text{reg}}$ is a regularizer that may include weight decay and spectral normalization.

Temporal coherence can be encouraged by penalizing inconsistencies of feature representations across consecutive frames. Let ϕ_k denote a feature map extracted from the video latent for frame k at some intermediate layer. A simple temporal loss can be defined as

$$\mathcal{L}_{\text{temp}} = \mathbb{E}\left[\sum_k \|\phi_{k+1} - \phi_k\|_2^2\right], [25] \tag{5.8}$$

possibly weighted by the mask to focus on regions where edits occur. This captures short-range temporal smoothness. Longer-range consistency can be addressed by considering multi-step differences or by tracking correspondences along estimated motion trajectories, but such enhancements increase computational complexity.

From an optimization standpoint, the training problem is a large-scale stochastic optimization in a high-dimensional parameter space. The gradient of the loss with respect to parameters θ is computed via backpropagation through the diffusion network and the various loss components. To maintain numerical stability and efficiency, gradient clipping and adaptive learning rate methods are often employed. The presence of region-aware weighting makes the gradient contributions from different spatial locations heterogeneous. Let g_i denote the gradient contribution from pixel i, and let w_i denote the corresponding mask-based weight. The total gradient can be expressed as

$$\nabla_\theta \mathcal{L}_{\text{denoise}} = \sum_i w_i g_i [26]. \tag{5.9}$$

If the weights vary significantly across pixels, the gradient may become dominated by a relatively small set of region pixels, potentially leading to unstable parameter updates. To mitigate this, one can normalize the weights or introduce a per-sample normalization factor

$$\tilde{w}_i = \frac{w_i}{\sum_j w_j}, \tag{5.10}$$

which ensures that the total amplitude of the weighted gradient remains bounded.

Sampling from a trained region-of-interest aware diffusion model involves integrating the reverse dynamics under the influence of the mask and the condition. In discrete time, the reverse sampler proceeds from $x_K \sim \mathcal{N}(0, I)$ down to x_0^* via a sequence of updates. When using a noise-prediction formulation, the update at step t can be written as

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} (x_t - \beta_t \epsilon_{\mathrm{roi}}(x_t, M, c, t)) + \sigma_t z_t, [27] \tag{5.11}$$

where β_t and σ_t are schedule-dependent coefficients. This is an explicit integration scheme whose stability is tied to the spectrum of the Jacobian of the denoiser. To improve stability and sample quality, higher-order numerical schemes can be employed. For example, a two-step method that approximates the drift at intermediate points can reduce local truncation error. Consider an approximation of the continuous-time reverse dynamics

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -b_t \epsilon_{\mathrm{roi}}(x, M, c, t), \tag{5.12}$$

where b_t is a scalar function. A second-order Runge–Kutta method can be implemented by computing

$$k_1 = -b_t \epsilon_{\mathrm{roi}}(x_t, M, c, t), \tag{5.13}$$

$$\tilde{x} = x_t + \Delta t\, k_1, \tag{5.14}$$

$$[28] k_2 = -b_{t-\Delta t} \epsilon_{\mathrm{roi}}(\tilde{x}, M, c, t - \Delta t), \tag{5.15}$$

$$x_{t-\Delta t} = x_t + \frac{\Delta t}{2}(k_1 + k_2), \tag{5.16}$$

with appropriate noise terms added to match the stochastic dynamics. Such higher-order integrators can reduce the number of function evaluations required for a given level of sample quality, which is particularly relevant when operating on large video tensors.

Discretization choices also influence how region-aware behavior manifests in the final video. If the timestep schedule allocates many steps at high noise levels and fewer at low noise levels, the model may have more opportunity to adjust global structure early in the trajectory but less opportunity for fine local refinements near the end. For region-of-interest editing, it may be beneficial to skew the schedule or adaptively adjust guidance strength as a function of time. For instance, one can define a time-dependent guidance field

$$g_t = \gamma_t M, \tag{5.17}$$

where γ_t is a scalar that increases as t approaches zero [29]. In early steps with large noise, guidance is weaker, allowing the reverse process to establish a coherent global structure, while in later steps guidance is stronger, focusing on refining the edited region. This strategy reflects the intuition that coarse structure should be formed before fine-grained edits are enforced.

In scenarios where the original video x_0 is available as an input to the editing process, one can initialize the reverse trajectory from a noisy version of x_0 rather than from pure Gaussian noise. This can be implemented by sampling a noise vector ϵ and constructing

$$x_T = \sqrt{\alpha_T} x_0 + \sqrt{1 - \alpha_T} \epsilon. \tag{5.18}$$

The reverse process then denoises x_T under the influence of the region-aware condition, leading to an edited version x_0^* that remains anchored to the original video. This initialization strategy often improves background preservation, since the model starts from a point already close to the original content, and the mask-guided denoising primarily alters the region of interest. However, it also increases the risk of residual artifacts if the denoiser fails to remove enough noise in the region while preserving the background.

6. Experimental Evaluation and Analysis

To evaluate region-of-interest aware diffusion models for controllable video editing, one can consider synthetic and real-world editing tasks that stress different aspects of controllability, identity preservation, and temporal coherence. In synthetic setups, masks and target edits can be derived from known transformations applied to original videos. For example, objects can be replaced with different categories, colors can be shifted, or textures can be altered, with corresponding ground truth edited videos generated through deterministic rendering or compositing pipelines. In real-world scenarios, masks may be obtained from segmentation algorithms or manual annotations, and target styles or attributes are specified via textual prompts or reference images [30].

A useful class of tasks involves object-centric edits in which a foreground object is replaced or modified while the background remains unchanged. Videos containing people, vehicles, or animals provide natural testbeds, as segmentation models can produce approximate foreground masks. In such tasks, region-of-interest aware diffusion models can be compared against baseline methods that use global conditioning without explicit mask integration. Quantitative metrics can include reconstruction error outside the region, measured as the average ℓ_1 or ℓ_2 difference between the original and edited videos in the background, as well as perceptual distances computed using pre-trained feature extractors. Within the region, similarity to a desired attribute or style can be approximated using classifier-based scores or feature distances relative to reference exemplars.

Temporal coherence can be analyzed by measuring frame-to-frame differences both in pixel space and in feature space. One can compute, for each video, the average temporal gradient magnitude

$$G_{\text{temp}} = \frac{1}{T-1} \sum_{k=1}^{T-1} \left\| x_0^*(k+1) - x_0^*(k) \right\|_2^2, \tag{6.1}$$

and compare it to the corresponding quantity for the original video. Values that are excessively large may indicate flickering or instability, whereas values that are unusually small could signal oversmoothing [31]. By computing G_{temp} separately for region and background, it is possible to study how region-aware guidance affects dynamic behavior in different parts of the frame. For example, a well-behaved edit may show increased temporal variation in the region when new motion patterns are introduced, while maintaining similar temporal gradients in the background.

Boundary quality at the interface of edited and preserved regions can be studied using mask-based band analysis. Define a boundary band by taking the set of pixels within a small distance from ∂M. Errors and temporal inconsistencies can be computed specifically within this band. Let B denote the index set of boundary pixels. A boundary distortion metric may be defined as

$$E_{\text{bound}} = \frac{1}{|B|} \sum_{i \in B} \left\| x_0^*(i) - x_0(i) \right\|_2^2. \tag{6.2}$$

Low values of this metric suggest that the transition between edited and non-edited regions is visually coherent and does not introduce noticeable artifacts [32]. When comparing region-aware models to baselines that ignore masks, one can analyze whether explicit mask conditioning reduces boundary distortions and yields smoother compositing.

From a probabilistic perspective, the effect of region-aware conditioning can be inferred by examining the distribution of latent codes at various diffusion steps. Consider extracting intermediate representations z_t from a fixed layer of the network during sampling. These representations can be decomposed into region and background components using the mask. One can then compute empirical covariance matrices for both components,

$$\Sigma_{\text{roi}}(t) = \mathbb{E}\left[(z_t^{\text{roi}} - \mu_{\text{roi}}(t))(z_t^{\text{roi}} - \mu_{\text{roi}}(t))^\top\right], \tag{6.3}$$

$$\Sigma_{\text{bg}}(t) = \mathbb{E}\left[(z_t^{\text{bg}} - \mu_{\text{bg}}(t))(z_t^{\text{bg}} - \mu_{\text{bg}}(t))^\top\right], \tag{6.4}$$

where $\mu_{\text{roi}}(t)$ and $\mu_{\text{bg}}(t)$ are mean vectors. Differences in the eigenvalue spectra of these covariance matrices across time steps can reveal how the model allocates representational capacity to the region versus the background [33]. For instance, a larger spread of eigenvalues for $\Sigma_{\text{roi}}(t)$ may indicate that the model represents a richer set of variations within the region, consistent with diverse possible edits, while a more concentrated spectrum for $\Sigma_{\text{bg}}(t)$ may reflect stronger regularization towards preserving original content.

Another aspect of analysis involves studying the sensitivity of the edited video with respect to perturbations in the mask. Given an input video and a condition, one can slightly dilate or erode the mask and observe how the output changes. A local sensitivity measure can be defined using finite differences. For a small perturbation δM, the change in the output can be approximated as

$$S = \frac{1}{\|\delta M\|_2} \left\|x_0^*(M + \delta M) - x_0^*(M)\right\|_2. \tag{6.5}$$

Low sensitivity away from the mask boundary suggests that the model is robust to small imperfections in the mask, which is desirable when masks are obtained from noisy segmentation algorithms. Higher sensitivity near the boundary may be acceptable, as mask edits in that region explicitly change which pixels are designated for editing.

Comparisons with baseline models that do not incorporate region-aware mechanisms can clarify the value of explicit region conditioning. For example, a baseline diffusion model trained with global conditioning can be adapted to approximate region-specific edits by blending its output with the original video using the mask. That is, one can generate a globally edited video y_0 and then form [34]

$$x_0^{\text{blend}} = M \odot y_0 + (1 - M) \odot x_0. \tag{6.6}$$

While this strategy ensures perfect background preservation, it may lead to inconsistencies at the boundaries, since the network did not account for the need to align the edited content with the original background. Artifacts such as halos, ghosts, or geometry misalignments may arise. In contrast, a region-of-interest aware diffusion model integrates the mask during denoising, enabling the network to adjust both edited content and its interface with the background jointly. Evaluating these two approaches using boundary metrics and perceptual scores can highlight the benefits of the integrated region-aware formulation.

User-controllable parameters play a significant role in practical editing. Besides the mask, parameters such as guidance strength, temporal consistency weights, and region expansion factors allow users to navigate the trade-offs between edit intensity and preservation of the original video. One can analyze the effect of varying these parameters by sampling multiple edited videos for the same input and condition. For each parameter configuration, metrics such as background error, region attribute alignment, and temporal gradient statistics can be recorded [35]. The resulting curves provide insight into how the region-aware diffusion model responds to control parameters and whether its behavior matches user expectations. For instance, increasing a guidance scale might produce stronger attribute changes but also increase the risk of overshooting or introducing artifacts.

Finally, computational considerations must be assessed. Operating on full-resolution video volumes is expensive, and region-aware conditioning introduces additional overhead in the form of mask processing and modified attention computations. The effective computational cost depends on the size of the region relative to the entire frame. If regions are typically small, one could consider hybrid approaches that restrict certain expensive operations, such as high-resolution temporal attention, to the region and its immediate vicinity, while applying cheaper operations elsewhere. Theoretical analysis of algorithmic complexity can quantify the dependence of runtime and memory on video length, resolution, and region size, and empirical benchmarks can validate whether the proposed designs are feasible for interactive editing workflows.

7. Conclusion

This paper has examined region-of-interest aware diffusion models for controllable video editing, focusing on how spatially localized masks can be integrated into the denoising process to achieve precise edits while preserving global coherence. By representing videos as high-dimensional tensors and masks as operators acting on these tensors, the analysis has connected region-aware conditioning to non-isotropic noise removal and spatially varying score fields. Several mechanisms for incorporating mask information into diffusion models have been considered, including masked feature decomposition, modulation of self-attention and cross-attention, and spatially varying guidance schemes derived from classifier-free guidance [36]. These mechanisms enable the model to treat edited and preserved regions differently in a unified generative framework.

The study has also discussed how spatiotemporal coherence can be promoted by combining region-aware denoising with temporal smoothness objectives and boundary-aware mask smoothing. The view of diffusion sampling as numerical integration of a reverse-time dynamical system has provided a lens for understanding how mask-based weighting affects the stability and contraction properties of the reverse process. Higher-order integration schemes and time-dependent guidance schedules have been highlighted as tools to balance global structure formation and local edit refinement. Training objectives with separate losses for region and background components have been proposed to manage the trade-off between edit fidelity and identity preservation, while considerations of gradient heterogeneity and normalization address numerical aspects of optimization.

Experimental analyses, in principle, can assess the behavior of region-of-interest aware diffusion models on tasks such as object replacement, localized stylization, and attribute modification, using metrics for background preservation, region attribute alignment, temporal gradients, and boundary quality. Comparisons with baseline methods that lack explicit region-aware conditioning, including simple blending of globally edited outputs with original videos, can clarify the practical advantages of integrating masks within the diffusion process. Sensitivity studies with respect to perturbations of the mask and control parameters such as guidance strength further characterize the robustness and controllability of the approach. Region-of-interest aware diffusion models provide a flexible framework for aligning generative video models with the needs of practical editing workflows, in which localized control, temporal stability, and background preservation are central considerations. The formulation developed here suggests several directions for further investigation, including mask-dependent architecture adaptations, dynamic mask evolution during sampling, and more advanced temporal modeling that exploits motion correspondences. As diffusion-based methods continue to evolve for video applications, incorporating structured spatial and temporal constraints through region-aware mechanisms may remain an important component of controllable video editing systems [37].

References

[1] F. Zhong, X. Qin, and Q. Peng, "Robust image segmentation against complex color distribution," *The Visual Computer*, vol. 27, pp. 707–716, 4 2011.

[2] J. Salau and J. Krieter, “Instance segmentation with mask r-cnn applied to loose-housed dairy cows in a multi-camera setting.,” *Animals : an open access journal from MDPI*, vol. 10, pp. 2402–, 12 2020.

[3] F. S. C. K and S. M. Kuriakose, “A review of super resolution and tumor detection techniques in medical imaging,” *International Journal of Trend in Scientific Research and Development*, vol. Volume-3, pp. 1785–1787, 4 2019.

[4] B. Zhu, C. Zhang, Y. Sui, and L. Li, “Facemotionpreserve: a generative approach for facial de-identification and medical information preservation.,” *Scientific reports*, vol. 14, pp. 17275–, 7 2024.

[5] P. Mora, C. Garcia, E. Ivorra, M. Ortega, and M. L. Alcañiz, “Virtual experience toolkit: An end-to-end automated 3d scene virtualization framework implementing computer vision techniques.,” *Sensors (Basel, Switzerland)*, vol. 24, pp. 3837–3837, 6 2024.

[6] F. Odone, A. Fusiello, and E. Trucco, “Layered representation of a video shot with mosaicing,” *Pattern Analysis & Applications*, vol. 5, pp. 296–305, 8 2002.

[7] S. S. Harsha, D. Agarwal, A. Revanur, and S. Agrawal, “Digital video editing based on a target digital image,” Aug. 21 2025. US Patent App. 18/583,067.

[8] M. Alahmadi, A. Khormi, B. Parajuli, J. Hassel, S. Haiduc, and P. Kumar, “Code localization in programming screencasts,” *Empirical Software Engineering*, vol. 25, pp. 1536–1572, 1 2020.

[9] S.-W. Jang and S.-H. Lee, “Robust blocking of human faces with personal information using artificial deep neural computing,” *Sustainability*, vol. 12, pp. 2373–, 3 2020.

[10] Y. Ma, J. Potappel, M. A. I. Schutyser, R. M. Boom, and L. Zhang, “Quantitative analysis of 3d food printing layer extrusion accuracy: Contextualizing automated image analysis with human evaluations: Quantifying 3d food printing accuracy.,” *Current research in food science*, vol. 6, pp. 100511–100511, 5 2023.

[11] W. Q. Yan, M. S. Kankanhalli, and J. Wang, “Analogies based video editing,” *Multimedia Systems*, vol. 11, pp. 3–18, 7 2005.

[12] L. min Xia, W. ting Guo, and W. Hao, “Interaction behavior recognition from multiple views,” *Journal of Central South University*, vol. 27, pp. 101–113, 1 2020.

[13] Z. Wu, T. Pei, Z. Bao, S. T. Ng, G. Lu, and K. Chen, “Utilizing intelligent technologies in construction and demolition waste management: From a systematic review to an implementation framework,” *Frontiers of Engineering Management*, vol. 12, pp. 1–23, 6 2024.

[14] S. S. Harsha, A. Revanur, D. Agarwal, and S. 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*, pp. 7559–7568, 2024.

[15] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Mégret, J. Pinquier, R. André-Obrecht, Y. Gaëstel, and J.-F. Dartigues, “Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia,” *Multimedia Tools and Applications*, vol. 69, pp. 743–771, 6 2012.

[16] S. C. Thomopoulos, *SYNTHESIS: A Platform of Platforms for Integrated Management, Curation, and Visualization of Digital Cultural Experiences through VR and AR Technologies*. IntechOpen, 9 2022.

[17] A. Nadeem, A. Jalal, and K. Kim, “Accurate physical activity recognition using multidimensional features and markov model for smart health fitness,” *Symmetry*, vol. 12, pp. 1766–, 10 2020.

[18] X. Jiang, Y. Du, and Y. Zheng, “Evaluation of physical education teaching effect using random forest model under artificial intelligence.,” *Heliyon*, vol. 10, pp. e23576–e23576, 12 2023.

[19] J. Tomás, A. Rego, S. Viciano-Tudela, and J. Lloret, “Incorrect facemask-wearing detection using convolutional neural networks with transfer learning,” *Healthcare (Basel, Switzerland)*, vol. 9, pp. 1050–, 8 2021.

[20] A. Hershkovitz, S. Knight, S. Dawson, J. Jovanovic, and D. Gašević, “About "learning" and "analytics",” *Journal of Learning Analytics*, vol. 3, pp. 1–5, 9 2016.

[21] A. Mankotia and M. Garg, “Real-time person segmentation – based onbody pix,” *International Journal for Modern Trends in Science and Technology*, vol. 6, pp. 1–7, 12 2020.

[22] C. M. Vasile and X. Iriart, “Embracing ai: The imperative tool for echo labs to stay ahead of the curve.,” *Diagnostics (Basel, Switzerland)*, vol. 13, pp. 3137–3137, 10 2023.

[23] Z. Yi, Q. Tang, V. S. R. Srinivasan, and Z. Xu, “Acm multimedia - animating through warping: An efficient method for high-quality facial expression animation,” in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1459–1468, ACM, 10 2020.

[24] C. Shi, J. He, S. Pundlik, X. Zhou, N. Wu, and G. Luo, “Low-cost real-time vlsi system for high-accuracy optical flow estimation using biological motion features and random forests,” *Science China Information Sciences*, vol. 66, 2 2023.

[25] Q. Wang, Z. Zhang, Q. Chen, J. Zhang, and S. Kang, “Lightweight transmission line fault detection method based on leaner yolov7-tiny.,” *Sensors (Basel, Switzerland)*, vol. 24, pp. 565–565, 1 2024.

[26] K. Kim, I. M. Alshenaifi, S. Ramachandran, J. Kim, T. Zia, and A. Almorjan, “Cybersecurity and cyber forensics for smart cities: A comprehensive literature review and survey.,” *Sensors (Basel, Switzerland)*, vol. 23, pp. 3681–3681, 4 2023.

[27] C. Zhang and Z. Liu, “Prior-free dependent motion segmentation using helmholtz-hodge decomposition based object-motion oriented map,” *Journal of Computer Science and Technology*, vol. 32, pp. 520–535, 5 2017.

[28] N. A. Shelke and S. S. Kasana, “A comprehensive survey on passive techniques for digital video forgery detection,” *Multimedia Tools and Applications*, vol. 80, pp. 6247–6310, 10 2020.

[29] L. Theodorou, D. Massiceti, L. M. Zintgraf, S. Stumpf, C. Morrison, E. Cutrell, M. T. Harris, and K. Hofmann, “Assets - disability-first dataset creation: Lessons from constructing a dataset for teachable object recognition with blind and low vision data collectors,” in *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–12, ACM, 10 2021.

[30] Y. Yang, “The potential energy of artificial intelligence technology in university education reform from the perspective of communication science,” *Mobile Information Systems*, vol. 2021, pp. 1–7, 9 2021.

[31] L. Sun, “Aerobics movement decomposition action teaching system based on intelligent vision sensor,” *Journal of Sensors*, vol. 2021, pp. 1–13, 11 2021.

[32] A. Chirtsov and V. Mikushev, “Digital support system for full-time and remote mass individualized education with elements of machine learning,” *SOCIETY. INTEGRATION. EDUCATION. Proceedings of the International Scientific Conference*, vol. 5, pp. 308–317, 5 2021.

[33] A. Arnold, “Marine surveillance with marine radar: theory, simulation, contribution to computer-aided ship wakes detection,” 2 2024.

[34] H. Sharma and N. Kanwal, “Video surveillance in smart cities: current status, challenges & future directions,” *Multimedia Tools and Applications*, vol. 84, pp. 15787–15832, 6 2024.

[35] M. Shin, W. Paik, B. C. Kim, and S. Hwang, “An iot platform with monitoring robot applying cnn-based context-aware learning.,” *Sensors (Basel, Switzerland)*, vol. 19, pp. 2525–, 6 2019.

[36] M. C. S. Manzanares, J. J. R. Diez, R. M. Sánchez, M. J. Z. Yáñez, and R. C. Menéndez, “Lifelong learning from sustainable education: An analysis with eye tracking and data mining techniques,” *Sustainability*, vol. 12, pp. 1970–, 3 2020.

[37] D. Kim, H. Kim, S. Lee, Q. Lee, M. Lee, J. Lee, and C. Jun, “Design and implementation of a two-wheeled vehicle safe driving evaluation system.,” *Sensors (Basel, Switzerland)*, vol. 24, pp. 4739–4739, 7 2024.