Academy for Global Science and Innovation (AGSI)

ARTICLE

Latent-Space Control in Diffusion-Based Editors: Balancing Global Style and Local Structure

Nguyen Khoa,[1] Tran Huy,[2] and Le Duyen[3]

[1]Saigon Pacific University, Department of Computer Science, Le Duan Boulevard, Ho Chi Minh City 700000, Vietnam
[2]Hanoi Metropolitan Institute of Technology, Department of Software Engineering, Kim Ma Street, Ba Dinh District, Hanoi 100000, Vietnam
[3]Mekong Frontier University, Department of Information Systems and E-Business, Hung Vuong Avenue, Can Tho 900000, Vietnam

Abstract

Diffusion-based image editors have become central tools for interactive visual manipulation, combining probabilistic generative models with flexible conditioning mechanisms. Many practical editing scenarios require simultaneous control of global image style and preservation of local scene structure, for example when re-rendering a photograph in a target artistic style while maintaining object identities, spatial layout, and fine geometric details. However, the latent spaces employed by diffusion models and their associated encoders often entangle style-related and structure-related factors, and naive conditioning or prompt-based guidance tends to either over-write local content or fail to propagate a coherent global appearance. This work studies latent-space control in diffusion-based editors with a particular focus on the trade-off between global style changes and local structure preservation. The paper develops a geometric perspective on latent diffusion spaces, where style manipulations are modeled as movements along a low-dimensional subspace while local structure is treated as a set of constraints induced by the image decoder and the diffusion denoising dynamics. Building on this perspective, we describe optimization-based formulations that explicitly balance style and structural losses, and we interpret several existing heuristic editing strategies as approximations to these formulations. We then outline practical algorithms that combine subspace projections, attention-weighted locality masks, and time-dependent control signals injected along the diffusion trajectory. Experimental scenarios are discussed qualitatively to illustrate how different design choices affect the balance between global coherence and local fidelity, and limitations and possible extensions of the approach are examined in detail.

1. Introduction

Diffusion-based generative models have become a prominent approach for high-quality image synthesis and editing, providing a flexible framework for conditioning on text, reference images, and other signals [1]. In editing applications, these models are typically deployed as diffusion-based editors that accept an input image, encode it to a latent representation, and then perform a controlled denoising process guided by user-specified constraints. A central challenge in this setting is achieving convincing global style transformations while maintaining the local structure and semantics of the original image. Examples include stylizing photographs into particular painting schools, harmonizing a foreground object into a newly generated background, or changing lighting and color mood without altering identities or geometrical relations.

Global style is often associated with broad-brush attributes such as color palettes, texture

Table 1. Key notation used in the latent-space control formulation for diffusion-based image editing.

Symbol	Meaning	Role in framework
z	Latent code of an input image	Starting point for editing in latent space
$x(z)$	Decoder output in pixel space	Rendered image corresponding to z
z_t	Latent state at diffusion time t	Controlled stochastic trajectory state
$J(z)$	Jacobian of decoder at z	Maps latent perturbations to pixel changes
$\phi(x)$	Style feature extractor	Defines global appearance statistics
$\psi(x)$	Structural feature extractor	Encodes local geometry and semantics
u_t	Control input in latent dynamics	Steers diffusion towards target edits

Table 2. Conceptual comparison between global style factors and local structure factors in diffusion-based editing.

Aspect	Global style	Local structure	Typical tension
Color statistics	Palette, tone, contrast	Local brightness consistency	Over-stylization can wash out details
Texture	Brush strokes, grain, noise	Material boundaries, fine edges	Strong textures may blur contours
Geometry	Perspective cues, shading	Object shape, pose, layout	Style may suggest altered geometry
Semantics	Domain (photo, comic, oil)	Identity, part relationships	Domain shifts can distort identity
Context	Scene mood, atmosphere	Background-foreground alignment	Style changes may break harmonization
Temporal/view consistency	Coherent look across frames	Stable motion and correspondences	Style drift vs. temporal stability

statistics, and lighting distributions that extend across the entire image. Local structure, in contrast, is tied to spatially localized features such as edges, contours, object shapes, and fine-grained geometry. In many practical tasks, users expect large stylistic shifts but low tolerance for structural distortions, for instance in portrait editing where facial geometry must remain stable under changes in rendering style. At the same time, strictly preserving structure can conflict with the generative models own internal prior over plausible images, especially when the target style implies structural motifs that differ from those in the original [2].

The latent representations used by diffusion-based editors, for example in latent diffusion models, complicate this tension. They are learned jointly with the denoising process and are optimized primarily for reconstruction and generative performance rather than disentanglement. As a result, straightforward editing strategies such as prompt substitution, classifier-free guidance, or latent interpolation often produce entangled changes in both style and structure. Attempts to mitigate this by masking, attention control, or feature blending partially address locality but do not directly reason about the underlying geometry of the latent space or the trade-off between competing objectives.

This paper develops a latent-space view of diffusion-based editing that explicitly models the balance between global style control and local structure preservation. The central idea is to interpret editing as a control problem in a high-dimensional stochastic dynamical system whose state evolves in latent space along the diffusion trajectory. Global style changes are expressed as movements along directions or subspaces that strongly affect appearance statistics, while local structure is encoded by constraints derived from the decoder Jacobian and spatial features. This geometric and optimization-based perspective provides a unifying language for comparing different editing heuristics and for designing new algorithms that manage the trade-off in a principled way [3].

The first part of the paper reviews diffusion-based editing and latent diffusion models, emphasizing how conditioning and noise schedules interact with latent encoders and decoders.

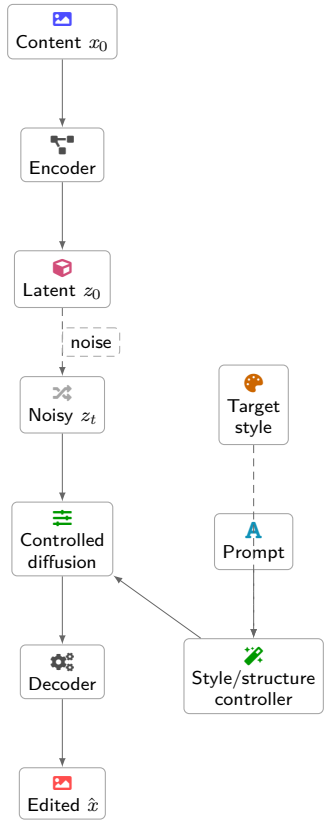


Figure 1. Overview of a diffusion-based image editor viewed in latent space. A content image is encoded into a latent representation, perturbed with noise, and then propagated through a controlled reverse diffusion process that is influenced by style references and prompts. The resulting latent is decoded back to pixels, yielding an edited image that reflects the interaction between global style guidance and structure-preserving control.

The second part introduces a decomposition of latent space into style- and structure-sensitive components, analyzes the role of the decoder Jacobian in relating latent perturbations to pixel-level changes, and proposes quantitative measures of structural distortion. Building on this, we formalize editing as constrained optimization in latent space and along the diffusion trajectory, where style objectives compete with structure-preserving regularizers. Practical algorithms are then described that realize these formulations by combining projected gradient steps, attention-weighted localization, and time-varying control strength.

The discussion includes qualitative experimental scenarios illustrating how the proposed formulations behave under different hyper-parameter choices, such as the dimensionality of the style subspace and the weight assigned to local structure. Rather than focusing on a specific benchmark, the emphasis is on analyzing characteristic failure modes and behaviors that occur across many editing tasks. The paper concludes by outlining limitations of linear subspace models of style, the sensitivity of structure preservation to encoder-decoder consistency, and potential directions for extending latent-space control to video, three-dimensional data, and more structured conditioning signals.

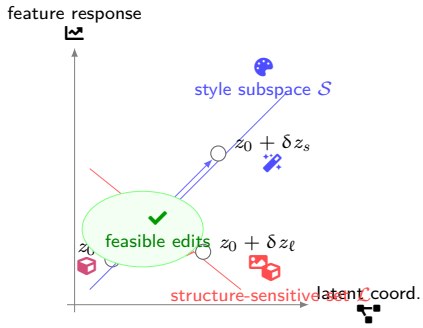


Figure 2. Schematic geometry of the latent space around an encoded image. Local linearization identifies directions that predominantly affect global style (style subspace $\mathcal{S}$) and directions that have strong impact on local structure. Feasible edits correspond to perturbations that move along style-related directions while staying within a region where structural distortions remain bounded, yielding a compromise between appearance change and content preservation.

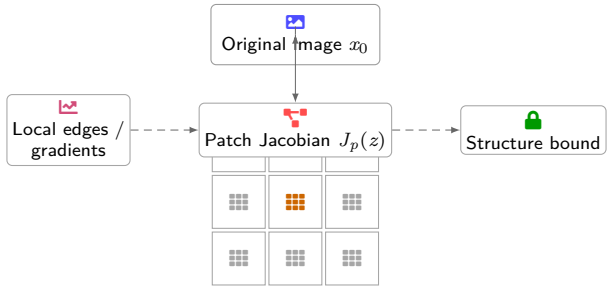


Figure 3. Patch-wise view of local structure preservation. The image is decomposed into small regions, each associated with a local Jacobian that describes how latent perturbations affect pixel intensities or features. Edge or gradient descriptors define structure-sensitive directions, and constraints on the corresponding Jacobian responses enforce that edits remain within acceptable tolerances for each patch.

2. Background on Diffusion-Based Editors

Diffusion models define a generative process by inverting a gradually corrupting forward process that adds noise to data. In the image domain, one considers a data distribution over images and defines a forward diffusion that transforms a clean image into a sequence of increasingly noisy versions. In discrete time formulations, this is often modeled as a Markov chain with time steps indexed by an integer variable. At each step, a small amount of Gaussian noise is added under a schedule that determines the variance at different times. The generative process then learns a parameterized reverse Markov chain that denoises from pure noise back to a sample from the data distribution.

Latent diffusion models modify this construction by first encoding images into a lower-dimensional latent space using an encoder network. The diffusion process operates in this latent space, and a decoder maps final latent samples back to pixel space. This design significantly reduces computational cost while maintaining perceptual quality, because the latent space is typically chosen so that local neighborhoods correspond to smooth variations in image space [4]. The diffusion network predicts noise or scores in latent space, and the loss function enforces consistency with the encoder-decoder reconstruction. For editing, an input image is

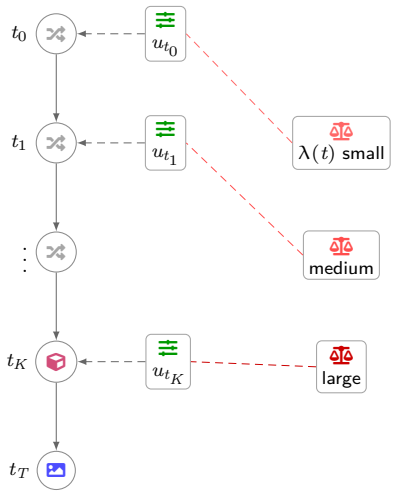


Figure 4. Controlled reverse diffusion as a time-indexed latent trajectory. At each diffusion time step t, a control term u_t is injected to steer the sampler toward the desired edit. A time-varying trade-off parameter $\lambda(t)$ modulates the relative strength of structure-preserving penalties, typically assigning weaker penalties at early, noisy steps and stronger penalties near the end of the trajectory when fine details are synthesized.

Table 3. Qualitative roles of different phases of the reverse diffusion trajectory in the proposed control scheme.

Diffusion stage	Noise regime	Control emphasis
Early steps	High noise, coarse content	Relaxed structure penalties, strong style exploration
Mid steps	Moderate noise, layout stabilized	Balanced style and structure objectives
Late steps	Low noise, detail synthesis	Strong structure preservation, cautious style updates
Final refinement	Very low noise	Mostly neutral directions, minor adjustments only
Ablative variant	Fixed control weights	Reveals trade-off without time-dependent scheduling

encoded to its latent representation, noise is added to obtain an intermediate diffusion state, and the reverse process is conditioned to realize the desired edit.

Conditioning mechanisms are central to diffusion-based editors. A common strategy is to use cross-attention to inject textual or visual conditioning into the denoiser. A text encoder produces embeddings of prompts, which are then used as keys and values in attention layers, while latent feature maps serve as queries. Visual reference images can be incorporated similarly, for example by encoding them into feature vectors or spatial feature maps that are fused with the latent representation. Classifier-free guidance and related techniques adjust the effective conditioning strength by combining conditional and unconditional noise predictions, often by extrapolating along directions in the space of predicted scores [5].

Editing operations can be framed as specific manipulations of conditioning signals, noise schedules, or latent trajectories. For instance, prompt-based editing may replace the original prompt with a new one, interpolate between prompts, or adjust guidance scales over time. More local edits may rely on spatial masks that restrict where conditional information is applied, or on modifications of attention maps to control which regions of the image respond to the new prompt. In many such designs, implicit assumptions are made about the relationship between the latent space geometry, the conditioning mechanisms, and the resulting changes

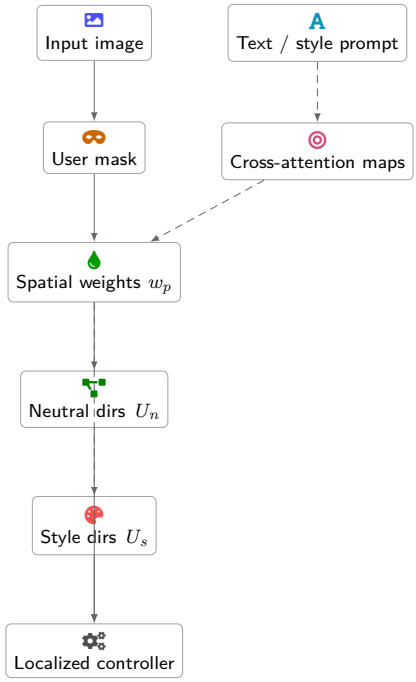


Figure 5. Local control of edits by combining user masks and attention-derived importance maps. Masked regions and attention responses are fused into spatial weights that shape the construction of structure-neutral directions U_n and style directions U_s. The resulting localized controller favors style changes inside the mask while enforcing stronger structure preservation elsewhere, leading to edits that respect both user-selected regions and model-driven locality.

in pixel space.

To discuss latent-space control, it is useful to formalize the latent representation. Let the encoder map an image to a latent tensor. The decoder approximates the inverse mapping. In practice, these are convolutional or transformer-based networks trained jointly with a perceptual loss and possibly adversarial components [6]. Although the latent space is not guaranteed to be globally linear, small neighborhoods can often be approximated as linear manifolds where latent perturbations map to pixel changes via a local Jacobian. This local linearity motivates the use of linear algebraic tools such as projections, subspace decompositions, and quadratic approximations to editing objectives [7].

Diffusion-based editors typically perform edits by initializing the reverse process from a noisy version of the encoded input. A chosen noise level determines how much of the original structure is preserved and how much freedom the model has to modify global attributes. High noise levels allow significant deviations from the original sample, enabling large style changes but risking structural drift. Low noise levels constrain the process tightly around the original image, preserving structure but sometimes limiting the achievable stylistic distance. This inherent trade-off is often tuned heuristically by choosing the starting time step of the reverse process.

Finally, evaluation of edits involves both global and local criteria [8]. Global style can be assessed using feature statistics extracted from deep networks, similarity of embeddings from multimodal encoders, or distribution-level measures of appearance. Local structure preserva-

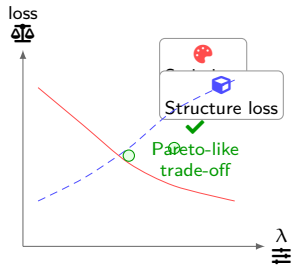


Figure 6. Illustration of the trade-off between global style objectives and local structure penalties as the weight λ is varied. Increasing λ typically suppresses structural distortion but also limits how far the edited image can move toward the target style. Intermediate values correspond to solutions near a Pareto frontier, which are of particular interest when visually balancing coherence of style changes with fidelity to the original content.

Table 4. Structural descriptors that can be used to define the local preservation loss in latent-space control.

Descriptor	Invariance	Main strengths	Main weaknesses
Pixel-wise differences	Low	Simple, cheap, sensitive to small changes	Overreacts to brightness and color shifts
Edge maps / gradients	Moderate to intensity shifts	Protects contours and silhouettes	Ignores high-level semantics
Deep recognition features	Semantic invariance	Preserves identity and object categories	May tolerate geometric drift
Keypoints / landmarks	Pose and part-aware	Enforces rigid geometric constraints	Requires detectors, sparse coverage
Segmentation features	Region-level invariance	Maintains object boundaries and masks	Coarse within-region structure

tion is reflected in metrics such as structural similarity, edge overlap, or keypoint alignment between the original and edited images. For latent-space analysis, one may also consider norms of latent differences projected onto subspaces capturing style or structure, linking geometric properties in latent space to perceptual outcomes. These considerations motivate the development of a more explicit mathematical treatment of latent-space control in diffusion-based editors.

3. Geometry of Latent-Space Control

We consider a latent space represented as a real vector space of dimension equal to the flattened size of the latent tensor produced by the encoder. Although the actual representation is spatially structured, it is convenient to treat the latent as a vector while keeping track of its spatial indexing separately. Let an input image be encoded as a latent vector denoted by a symbol that we view as an element of this high-dimensional space [9]. The decoder is a mapping from latent space to pixel space. For analysis, we regard this decoder as differentiable almost everywhere, so that a Jacobian matrix exists at points of interest.

Editing can be seen as choosing a perturbation of the original latent that leads, after decoding and optionally running the diffusion reverse process, to an image with modified global style and preserved local structure. From a geometric perspective, one would like to identify directions in latent space that primarily affect style, and directions that primarily affect structure. While exact disentanglement is not guaranteed, one can aim to construct approximate subspaces representing these roles. Let us denote a decomposition of the latent space into two linear subspaces that sum to the full space, without requiring orthogonality a priori. Vectors in one subspace are intended to capture global style variation, while vectors in the other subspace reflect structural variations that should be suppressed or constrained during

Table 5. Components of the optimization objective governing the controlled diffusion trajectory.

Objective term	Role	Effect when its weight increases
$L_{\text{sty}}(z_t)$	Aligns decoded image with target style	Stronger global style, risk of structural drift
$L_{\text{loc}}(z_t)$	Preserves local structure vs. original	Higher fidelity, potentially weaker stylization
$\|0u_t\|0_2^2$	Regularizes control magnitude	Smoother trajectories, fewer aggressive edits
λ	Trade-off between style and structure	Larger values favor structure over style
γ	Trade-off between control and losses	Limits overfitting to losses via extreme controls

Table 6. Representative qualitative experiments illustrating the trade-off between global style control and local structure preservation.

Scenario	Input content	Target change	Evaluation focus
Portrait stylization	Human face photos	Map to artistic domains (e.g., oil, comic)	Preserve identity and facial geometry
Street scene transfer	Urban photographs	Convert to specific painting styles	Maintain layout and object placements
Local object restyling	Single salient object	Change material or pattern only	Avoid halo artifacts at object boundaries
Background replacement	Fixed foreground subject	Generate new stylized background	Harmonize style while preserving subject
Material/cloth edit	Clothing or surface	Modify fabric type or texture	Keep pose and body shape intact

editing.

A simple way to formalize this idea uses linear projections [10]. Let two linear maps be defined such that their sum equals the identity operator on the latent space. For a given latent vector, we write

$$z_s = P_s z, \quad z_\ell = P_\ell z,$$

where the style component lies in the range of one projector and the local component lies in the range of the other. If the projectors satisfy idempotency and complementary properties, then every latent can be decomposed uniquely as a sum of a style part and a local part. In practice, these projectors are not known a priori and must be estimated from data or from the behavior of the decoder and diffusion dynamics [11].

One way to relate these subspaces to image properties is through the decoder Jacobian. For a latent vector, the Jacobian is a matrix that maps first-order latent perturbations to first-order changes in pixel space. Let us denote this Jacobian by a matrix. For a small latent displacement, the decoded image can be approximated as

$$x(z + \delta z) \approx x(z) + J(z)\,\delta z.$$

Here, first-order Taylor expansion captures how infinitesimal changes in latent coordinates affect the image [12]. If we restrict perturbations to the style subspace, we obtain changes of the form with the local component frozen. The resulting image variation then lies approximately in the column space of the product of the Jacobian with the style subspace. Similarly, restricting to the local subspace yields another family of variations.

Local structure preservation can be expressed as constraints on the Jacobian-propagated perturbations within spatial neighborhoods. Consider a set of spatial locations or pixels and define, for each location, a linear functional that extracts the intensity or a vector of features at that location. Combining these with the Jacobian yields localized sensitivity measures. For a patch-level projection, we can define a selector matrix that picks out a subset of pixel coordinates corresponding to a patch. The patch-level Jacobian is then the product of this selector with the full Jacobian [13]. For a perturbation in latent space, the corresponding change in that patch is approximated by this patch-level Jacobian acting on the perturbation.

Table 7. Latent subspace designs used to separate style-changing and structure-preserving directions in the editor.

Subspace	Construction approach	Controlled factors	Potential risks
Style subspace U_s	Leading directions of style-loss Hessian	Global color, texture, domain shift	Over-editing geometry and semantics
Neutral subspace U_n	Near-null directions of structural quadratic form	Small impact on local structure	Residual subtle distortions, halo artifacts
Structure-sensitive subspace	High-eigenvalue directions of structural form	Strong influence on geometry	Must be suppressed during late steps
Learned control directions	Trained on data for edits	Task-specific manipulation	Limited generalization to unseen styles
Random baseline directions	Unconstrained latent directions	Diagnostic comparison	Unpredictable style/structure entanglement

Table 8. Main limitations of the presented latent-space control framework and directions for further development.

Aspect	Current limitation	Possible extension
Linearity assumptions	Style modeled as low-dimensional linear subspace	Nonlinear manifolds or flow-based latent warps
Jacobian estimation	Expensive and noisy in high dimensions	Learned surrogates or amortized sensitivity models
Hyperparameter tuning	Manual schedules for λ and γ	Data-driven or perceptual-quality-based adaptation
Computational cost	Extra overhead per diffusion step	Low-rank updates, caching, and pruning of bases
Structural descriptors	Choice depends on task and dataset	Hybrid metrics combining edges and deep features
Multi-image consistency	Focus on single-image editing	Extending control to video and multi-view settings

To quantify structure preservation, we can insist that the component of the patch change orthogonal to some reference local structure representation remains small. For example, if edges or gradients of the original image define preferred directions of variation in a patch, then we can constrain the projection of the perturbation-induced change onto the orthogonal complement of these directions. This leads to a condition of the form

$$\left|0 R_p J_p(z)\,\delta z\right| 0_2[14] \le \epsilon_p,$$

where is a projector onto structure-orthogonal directions in patch and is a tolerance parameter. Aggregating across patches yields constraints that define a feasible set of latent perturbations preserving local structure within prescribed limits.

Global style, on the other hand, can be modeled by statistics aggregated over the whole image or over large regions. Suppose we define a feature extractor that maps images to a lower-dimensional feature vector capturing style-relevant statistics, such as channel-wise means, covariances, or deep activations pooled spatially. For a target style feature vector, a style loss can be defined as a function of the difference between the style features of the decoded image and the target. Linearizing around the original latent, we can approximate the change in style features as

$$\Delta s[15] \approx S(z)\,\delta z,$$

where is the Jacobian of the style feature extractor composed with the decoder. The style loss can then be approximated as a quadratic function of the latent perturbation, whose gradient is aligned with directions in latent space that most increase alignment with the target style.

From this viewpoint, the style subspace can be defined as the span of leading eigenvectors of the symmetric matrix obtained from the linearized style loss. The structure subspace can be associated with directions that minimally affect the structural constraints, for example those lying approximately in the null spaces of patch-level Jacobians after projection onto structure-relevant components. These constructions yield approximate style and structure subspaces that depend on the particular image and on the current latent position along the diffusion trajectory, reflecting the local geometry of the decoder and feature extractors [16].

Finally, the diffusion dynamics impose an additional layer of geometry. The reverse diffusion process can be expressed as a stochastic difference or differential equation in latent space,

driven by the learned score or noise prediction. Editing can be seen as adding a control term to this dynamical system, steering the trajectory towards latents that decode to images with the desired style while respecting structural constraints. The analysis of such controlled stochastic systems often relies on approximations of the drift and diffusion terms, as well as on local linearizations of the dynamics around a nominal trajectory. Within these approximations, the latent space geometry described above shapes the effective control directions and the trade-offs between global and local objectives.

4. Methods for Balancing Global Style and Local Structure

We now formalize editing in diffusion-based latent spaces as an optimization and control problem that explicitly trades off global style objectives against local structure preservation. Let an input image be encoded to a latent representation. Let there be a stochastic reverse diffusion trajectory in latent space, typically indexed by discrete time steps or continuous time in some interval [17]. In standard unconditional generation, this trajectory starts from a noise sample and evolves according to a learned denoising network. In editing mode, the trajectory is initialized from a noisy version of the encoded latent, and conditioning is used to guide the evolution.

From a control perspective, we view the latent trajectory as the state of a controlled dynamical system. In a continuous-time formulation, an approximate reverse stochastic differential equation in latent space can be written as

$$\mathrm{d}z_t = f_\theta(z_t, t, c)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t + u_t\,\mathrm{d}t,$$

where is the latent state at time, is the drift term determined by the learned score or noise prediction conditioned on a control signal, is a diffusion coefficient multiplying a Wiener process, and is an additional control term that we introduce for editing. In practice, time is discretized and the dynamics are implemented via a finite-step sampler, but the continuous formulation clarifies the role of control inputs.

Global style control is encoded through a loss function that measures the discrepancy between the current image decoded from and a target style. Let be the composition of the decoder with a style feature extractor [18]. We define style features and a target style vector, and a style loss of the form

$$L_{\mathrm{sty}}(z_t) = \ell_s\big(\phi(x(z_t)), s^\star\big)\,,$$

where is a differentiable function, often a squared norm or more complex distance in feature space. Local structure preservation is encoded through a loss that penalizes deviations of structural descriptors between the edited image and the original image [19]. Let be a structural feature extractor, possibly depending on spatial positions or patches, and denote the structural features of the original image by. Then we define

$$L_{\mathrm{loc}}(z_t) = \ell_\ell\big(\psi(x(z_t)), \psi(x_0)\big)\,,$$

where is a structural discrepancy measure, for instance a weighted sum of patchwise distances.

Editing is then posed as minimizing a cumulative objective over the diffusion trajectory, such as [20]

$$\mathcal{J}(u) = \mathbb{E}\left[\int_0^1 \Big(L_{\mathrm{sty}}(z_t) + \lambda\, L_{\mathrm{loc}}(z_t) + \gamma\,|0u_t|0_2^2\Big)\,\mathrm{d}t\right], [21]$$

where the expectation is with respect to the stochasticity of the diffusion process, is a trade-off parameter balancing style and local structure, and is a weight penalizing control effort

to avoid extreme manipulations. In practice, we approximate this objective using a single or small number of trajectories and discrete sums over time steps. The main design problem is to specify a parameterization of the control term that respects computational constraints and exploits the latent geometry.

To incorporate the style-structure decomposition, we parameterize control as taking values in a style subspace plus corrections in a structure-neutral subspace. Let be matrices whose columns form bases of approximate style and structure-neutral subspaces at time. We write the control as

$$u_t = U_s(t)\,\alpha_t + U_n(t)\,\beta_t, [22]$$

where are low-dimensional coefficient vectors. The style basis is constructed to approximate directions in latent space that produce strong variation of the style feature extractor, for example via local principal component analysis of gradients of the style loss. The neutral basis is chosen to approximate directions that minimally affect structural descriptors, for instance via approximate null spaces of patch-level Jacobians of the composition of the decoder with the structural feature extractor.

Given this parameterization, the gradients of the objective with respect to the control coefficients can be obtained by backpropagating through the sampled diffusion trajectory or by local approximations. A simple scheme performs greedy updates at each time step, treating as fixed and computing a control increment that locally minimizes a weighted sum of style and structural losses. In a discrete-time sampler with states, predicted noise, and step sizes, we can write an update of the form

$$z_{k+1} = z_k + \Delta_k(z_k, c) + U_s(k)\,\alpha_k[23] + U_n(k)\,\beta_k,$$

where is the base sampler update at step. The control coefficients are chosen, for instance, by one or a few iterations of a gradient-based update on a local objective

$$\mathcal{L}_k(\alpha_k, \beta_k) = L_{\text{sty}}(z_{k+1}) + \lambda_k L_{\text{loc}}(z_{k+1}) + \gamma_k \left|0\,U_s(k)\,\alpha_k[24] + U_n(k)\,\beta_k\right| 0_2^2,$$

with a time-dependent trade-off parameter. Early steps in the diffusion process can be assigned lower values, allowing larger structural deviations to establish global style, while later steps use higher values to lock in local structure.

Locality is introduced by modulating the contribution of different spatial positions in the latent tensor. Let the latent be reshaped into a spatial grid of latent vectors at each position. Local structure constraints can be applied patchwise or per position, and the neutral basis can be constructed to minimize structural change in selected regions. Attention maps derived from the diffusion models cross-attention layers provide spatial importance weights that indicate where conditioning signals are most active for given tokens [25]. These weights can be used to build masks that soften structural constraints in regions where modifications are desired, while enforcing stronger constraints elsewhere.

Formally, let be a diagonal matrix of spatial weights derived from attention or user-provided masks, with entries in the range between zero and one. The structural loss can be expressed as a weighted sum over spatial positions, and local neutral directions can be constructed by solving approximate eigenvalue problems for weighted Jacobian matrices. For example, one may define a regularized quadratic form

$$Q(\delta z) = \sum_p w_p \left|0[26]R_p J_p(z)\,\delta z\right| 0_2^2 + \eta|0\delta z|0_2^2,$$

where index patches or positions, are weights, and is a small regularization parameter. Directions with small values of relative to their norm form candidate neutral directions respecting

local structure. A small number of such directions can be approximated by iterative methods on the fly, or precomputed for representative configurations.

To keep computation tractable, further approximations are typically required. Instead of full Jacobians, one can use linear mappings induced by feature differences for small random perturbations of the latent, approximating local sensitivities via finite differences [27]. Rather than explicit eigenvector computations, one can maintain a small pool of candidate directions generated from gradients of structural and style losses, orthogonalizing and reweighting them to form bases. This yields adaptive subspaces that evolve along the diffusion trajectory as the latent state moves.

Finally, the hyper-parameters governing the balance between global style and local structure are scheduled over time. A common pattern is to start with more freedom for style control at high noise levels, where the model is reconstructing coarse structure, and then gradually increase structural regularization as the noise decreases and details are synthesized. The schedule may depend on the noise variance, on heuristics about the progress of denoising, or on online measurements of style and structural losses. The resulting algorithm produces, for each given input image and target style, a latent trajectory whose endpoint decodes to an image that reflects the negotiated compromise between global coherence and local fidelity.

5. Experimental Evaluation

To examine the behavior of the proposed latent-space control formulations, one can consider a range of qualitative experimental setups that emphasize different aspects of the global style versus local structure trade-off. Since the focus is on understanding characteristic patterns rather than reporting specific numerical benchmarks, the discussion here remains descriptive [28]. Nevertheless, the same ideas can be adapted to more formal evaluations using standard datasets and metrics.

A natural test scenario is style-preserving content editing, where an input photograph is stylized to resemble a particular artistic domain while preserving object identities, pose, and spatial layout. For example, one might use portraits, street scenes, or landscapes as content images and requests such as transforming the scene into watercolor, oil painting, or comic-book style as target styles. Baseline methods such as prompt replacement with fixed classifier-free guidance, or purely mask-based editing that restricts modifications to selected regions, provide reference behaviors. These baselines often exhibit either over-editing, where facial features are distorted, edges are blurred, or object shapes are altered, or under-editing, where stylistic changes remain limited to color shifts and weak texture overlays.

Applying the latent-space control approach with a style subspace and local structural constraints yields different trends as the trade-off parameter is varied. For small values, the algorithm behaves similarly to aggressive global style transfer: the latent trajectory explores directions of high style gradient, and the resulting images may approximate the visual statistics of the target domain but deviate substantially from the original structure [29]. As is increased, the structural loss begins to dominate in later diffusion steps, constraining the control directions to lie closer to the neutral subspace. This leads to images where contours, geometry, and object placements remain aligned with the original, while still showing global adjustments in color, tone, and texture that reflect the target style.

Qualitative inspection reveals that the time-dependent scheduling of is particularly important. If structural penalties are enforced strongly from early steps, the model sometimes fails to acquire distinctive style features and produces results that look like lightly filtered versions of

the original image. In contrast, allowing early steps to move more freely in style directions, then gradually restricting the control to structure-neutral components as noise decreases, leads to images that better reconcile style and content. This behavior aligns with the intuition that coarse-scale layout is largely determined in earlier stages of denoising, while finer details and textures are synthesized later.

Another class of experiments involves local edits such as changing the material of a single object, altering clothing style on a person, or modifying the rendering of the sky while keeping the foreground intact. Here, attention-derived masks and user-provided masks are used to modulate local structure penalties [30]. Regions outside the editing target are assigned higher structural weights, forcing the optimization to respect their geometry, while the target region is given lower weights that allow more freedom. In the latent-space framework, this is reflected both in the construction of neutral subspaces, which take into account local Jacobians weighted by patch importance, and in the structural loss term, which aggregates errors with spatial weights.

Comparisons with purely mask-based editing that only restricts where conditional information is applied show differences in failure modes. Mask-based approaches often suffer from boundary artifacts and inconsistent style at mask edges, because the latent updates do not explicitly account for how changes in the edited region propagate through the decoders receptive fields into neighboring pixels. Subspace-based control with structural penalties tends to produce smoother transitions, since candidate control directions that would significantly distort neighboring structure are downweighted or removed. However, when masks are very tight and the neutral subspace is constructed using approximate Jacobians, there can still be subtle halo artifacts, indicating the limits of local linear approximations.

An additional set of experiments investigates the effect of the dimensionality of the style subspace [31]. When the number of style directions is very small, edits are constrained to lie along a low-dimensional manifold of style changes, leading to stable but sometimes conservative transformations. As the dimensionality increases, the space of possible style manipulations expands, allowing more diverse and sometimes more faithful reconstructions of target styles, at the cost of greater risk of structural deviations. In practice, moderate dimensions often provide a compromise, especially when combined with time-dependent structural penalties that buffer the effect of high-dimensional style control at late diffusion steps.

Alternative choices for structural descriptors and losses also influence the observed behavior. Using simple pixel-wise differences or edge detectors provides a basic notion of structure but can be overly sensitive to brightness changes and minor stylistic variations. Incorporating deep features from networks trained for recognition or segmentation yields more semantically aware structure measures that distinguish between stylistic repainting and changes that alter object identity or shape. Latent-space control using such semantically informed structural losses tends to prioritize preservation of object-level attributes, sometimes at the expense of fine textural consistency with the target style.

Finally, ablations on the control parameterization itself illustrate the role of the neutral subspace [32]. If control is restricted to the style subspace without any neutral component, then when structural penalties become strong the optimizer may be forced into very small control magnitudes, limiting the achievable style adjustment. Introducing a neutral subspace allows the algorithm to reorient control directions to reduce structural cost while maintaining some style improvement. However, if the neutral subspace is poorly aligned with true structure-preserving directions, for instance due to inaccurate Jacobian approximations, the resulting edits can exhibit unexpected artifacts. This underscores the importance of accurate local

geometry estimation and suggests potential benefits from more sophisticated approximations or learned mappings that directly predict structure-preserving latent directions.

6. Discussion and Limitations

The latent-space control perspective developed here frames diffusion-based editing as a controlled stochastic dynamical system where style and structure objectives compete. This framing has several conceptual implications. First, it emphasizes that many editing heuristics, such as gradually adjusting classifier-free guidance or modifying attention weights, can be interpreted as particular choices of control inputs in latent space, albeit ones that are implicitly defined by network architectures rather than explicit optimization. Viewing these operations as control policies clarifies the roles of different components and enables systematic exploration of alternative strategies [33].

Second, the decomposition of latent space into style and structure-related subspaces highlights the importance of local geometry induced by the encoder-decoder pair and by feature extractors. In idealized cases where style and structure factors are aligned with linear subspaces, the proposed methods can achieve clean separation between global appearance changes and local geometric constraints. In realistic settings, however, the latent representation is only approximately linear over the region explored during editing, and the subspaces derived from local Jacobians and gradients are valid only within limited neighborhoods. As the diffusion trajectory moves in latent space, the local linear model changes, and subspaces must adapt accordingly. This introduces both additional computational cost and sensitivity to hyper-parameters controlling how rapidly the subspaces are updated.

Third, the reliance on local Jacobian information raises questions about robustness to encoder-decoder mismatches and numerical approximations. Exact Jacobians of high-dimensional decoders are expensive to compute, and in practice one relies on backpropagation through small sets of scalar functionals or finite-difference schemes that approximate directional derivatives [34]. These approximations can be noisy, especially in regions of latent space where the decoder behaves nonlinearly or where gradients are small. When structural constraints are enforced based on such noisy estimates, the resulting neutral subspaces may include directions that subtly alter structure or exclude directions that would be safe, leading to occasional artifacts or overconstrained edits [35].

Moreover, the structural descriptors chosen to define local losses influence which aspects of the image are prioritized. Simple edge-based descriptors may protect contours but allow changes to textures that affect perceived material properties, while deep-feature-based descriptors may preserve semantic content but permit geometric distortions that remain under the radar of feature differences. Selecting and combining structural descriptors therefore becomes an important modeling choice, and different applications may require different definitions of local structure. For instance, medical imaging may prioritize anatomical boundaries and intensities, while artistic style transfer may care more about preserving recognizable objects than exact geometrical fidelity.

The control objective balancing style and structural losses also brings to the foreground the issue of selecting trade-off weights and time schedules. In the absence of task-specific validation metrics, these hyper-parameters are often tuned empirically and may vary with the type of input image and target style [36]. A single global setting of the trade-off parameter may not be ideal across diverse scenes and styles, suggesting the potential usefulness of adaptive schemes that adjust the balance based on online measurements of editing progress or on learned predictors of perceptual quality.

From a computational standpoint, constructing and updating subspaces, computing gradients of style and structural losses, and performing local optimization at each diffusion step all add overhead compared to simpler editing strategies. While some of this overhead can be mitigated through dimensionality reduction, caching of Jacobian information, and approximate solvers, there remains a cost that may limit real-time interactivity for large models or high-resolution images. This constraint may be particularly relevant for applications that require many trial edits, such as iterative design workflows.

There are also conceptual limitations associated with modeling style primarily as a low-dimensional linear subspace of latent space. While this assumption is convenient and aligns with interpretations of style as a global set of features or statistics, real-world styles can involve complex nonlinear changes, including spatially varying patterns and high-order correlations across image regions. Capturing such styles may require more expressive models, such as nonlinear manifolds in latent space or conditional flows that warp latent neighborhoods in style-specific ways [37]. Extending the control framework to these richer geometric structures remains an open direction.

Finally, the analysis has focused on single-image editing, whereas many applications involve temporal or spatial consistency across multiple images, such as video editing or view-consistent editing of three-dimensional scenes. In such settings, latent-space control must account not only for local spatial structure within individual frames but also for coherence over time or across viewpoints. The corresponding structural descriptors become higher dimensional, involving trajectories of features or multi-view correspondences, and the diffusion process may be defined over sequences or volumetric representations. Adapting the proposed methods to these more complex domains is nontrivial and would likely require additional modeling of temporal and geometric constraints.

7. Conclusion

This paper has examined latent-space control in diffusion-based editors with the goal of balancing global style manipulation and local structure preservation. Starting from the observation that naive conditioning and prompt-based editing often entangle these aspects, we developed a geometric view of latent diffusion spaces that distinguishes approximate style and structure subspaces and relates them to the decoder Jacobian and feature-based descriptors of images. Within this view, editing is formulated as a control problem for the diffusion reverse process, with a latent control term that is optimized to minimize a cumulative objective combining style and structural losses [38].

We described how linear-algebraic tools and local linearizations can be employed to construct style subspaces aligned with gradients of style objectives and structure-neutral subspaces derived from patch-level Jacobians and structural descriptors. Based on these constructions, we proposed parameterizations of control inputs that operate within these subspaces and are updated along the diffusion trajectory according to a time-dependent schedule that reflects the evolving balance between global appearance changes and structural fidelity. Attention-derived masks and spatial weighting were used to localize structure constraints, enabling both global and region-specific edits within the same framework.

Qualitative experimental considerations illustrated how the proposed approach can modulate the strength of style transfer while controlling structural deviations, how time-dependent trade-off schedules influence outcomes, and how design choices such as the dimensionality of style subspaces and the choice of structural descriptors affect behavior. The discussion highlighted limitations arising from approximate geometry, hyper-parameter sensitivity, and

computational cost, as well as conceptual issues related to the linear modeling of style and the extension to temporally or spatially coherent multi-image settings. Treating diffusion-based editing as latent-space control offers a structured way to reason about the interactions between conditioning, latent geometry, and perceptual properties of edited images. While the specific formulations explored here rely on simplifying assumptions and local approximations, they provide a foundation upon which more sophisticated control strategies, richer geometric models, and application-specific structural constraints can be built. Future work may investigate learned predictors of structure-preserving directions, data-driven adaptation of trade-off schedules, and extensions of the framework to video, three-dimensional content, and multimodal editing scenarios where multiple sources of style and structure must be reconciled simultaneously [39].

References

[1] James Park. *Symmetry-Adapted Machine Learning for Information Security*. MDPI, Aug. 7, 2020. DOI: 10.3390/books978-3-03936-643-9.

[2] Oussama Zayene, Sameh Masmoudi Touj, Jean Hennebert, Rolf Ingold, and Najoua Essoukri Ben Amara. "Open Datasets and Tools for Arabic Text Detection and Recognition in News Video Frames". In: *Journal of Imaging* 4.2 (Jan. 31, 2018), pp. 32–. DOI: 10.3390/jimaging4020032.

[3] Joanna Isabelle Olszewska. "Automated Face Recognition: Challenges and Solutions". In: InTech, Dec. 14, 2016, pp. 59–79. DOI: 10.5772/66013.

[4] Birgit Schauer. *Deliverable 3.2: Report on technical implementation of the platform*. Feb. 2, 2023. DOI: 10.5281/zenodo.7598484.

[5] Kimiaki Shirahama and Marcin Grzegorzek. "Towards large-scale multimedia retrieval enriched by knowledge about human interpretation". In: *Multimedia Tools and Applications* 75.1 (Oct. 5, 2014), pp. 297–331. DOI: 10.1007/s11042-014-2292-8.

[6] Qingbo Yan. "Design of Teaching Video Resource Management System in Colleges and Universities based on Microtechnology". In: *Security and Communication Networks* 2021 (May 19, 2021), pp. 1–11. DOI: 10.1155/2021/9914048.

[7] Ambareesh Revanur, Debraj Debashish Basu, Shradha Agrawal, Dhwanit Agarwal, and Deepak Pai. *MASK CONDITIONED IMAGE TRANSFORMATION BASED ON A TEXT PROMPT*. US Patent App. 18/319,808. Nov. 2024.

[8] Shin-Hung Pan and Shu-Ching Wang. "Identifying Vehicles Dynamically on Freeway CCTV Images through the YOLO Deep Learning Model". In: *Sensors and Materials* 33.5 (May 6, 2021), pp. 1517–. DOI: 10.18494/sam.2021.3236.

[9] Edirlei Soares de Lima, Bruno Feijó, and Antonio L. Furtado. "Video-based interactive storytelling using real-time video compositing techniques". In: *Multimedia Tools and Applications* 77.2 (Feb. 2, 2017), pp. 2333–2357. DOI: 10.1007/s11042-017-4423-5.

[10] Jian Tang, Tianzheng Wang, Heng Xia, and Canlin Cui. *An Overview of Artificial Intelligence Application for Optimal Control of Municipal Solid Waste Incineration Process*. Jan. 24, 2024. DOI: 10.20944/preprints202401.1718.v1.

[11] Ethan Kyzivat and Laurence Smith. *Contemporary and historical detection of small lakes using cross-sensor super resolution Landsat imagery*. Feb. 18, 2024. DOI: 10.31223/x5ms9b.

[12] Osman Tugay Baaran, Yekta Said Can, Elisabeth André, and Cem Ersoy. "Relieving the burden of intensive labeling for stress monitoring in the wild by using semi-supervised learning." In: *Frontiers in psychology* 14 (Jan. 5, 2024), pp. 1293513–. DOI: 10.3389/fpsyg.2023.1293513.

[13] Margaret Allan. "The introduction of a new technology within one discipline". In: *Australasian Journal of Educational Technology* 3.2 (Dec. 1, 1987). DOI: 10.14742/ajet.2381.

[14] Alexander E. Novikov. *SKYNET 2023 Conception of the Artificial Super Intelligence Project. A System Approach. ver. 3.* Dec. 21, 2023. DOI: 10.31219/osf.io/kqt9p.

[15] Soma Tomihara, Yoshitaka Oka, and Shinji Kanda. "Establishment of open-source semi-automated behavioral analysis system and quantification of the difference of sexual motivation between laboratory and wild strains." In: *Scientific reports* 11.1 (May 25, 2021), pp. 10894–10894. DOI: 10.1038/s41598-021-90225-3.

[16] Xiaofeng Chang, Jintao Yu, Ying Gao, Hongchen Ding, Yulong Liu, and Huaming Yu. "Classification of Sailboat Tell Tail Based on Deep Learning". In: *Journal of Ocean University of China* 23.3 (May 20, 2024), pp. 710–720. DOI: 10.1007/s11802-024-5695-9.

[17] Bushra Yasmeen, Haslina Arshad, and Hameedur Rahman. "SUSPICIOUS ACTIVITY DETECTION USING CCTV SURVEILLANCE VIDEO". In: *Journal of Information System and Technology Management* 6.22 (Sept. 1, 2021), pp. 60–70. DOI: 10.35631/jistm.622006.

[18] Natalia Díaz-Rodríguez and Galena Pisoni. "UMAP (Adjunct Publication) - Accessible Cultural Heritage through Explainable Artificial Intelligence". In: *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, July 13, 2020, pp. 317–324. DOI: 10.1145/3386392.3399276.

[19] Yuxin Peng, Wenwu Zhu, Yao Zhao, Chang-sheng Xu, Qingming Huang, Han-qing Lu, Qinghua Zheng, Tiejun Huang, and Wen Gao. "Cross-media analysis and reasoning: advances and directions". In: *Frontiers of Information Technology & Electronic Engineering* 18.1 (Feb. 4, 2017), pp. 44–57. DOI: 10.1631/fitee.1601787.

[20] Dohun Kim, Heegwang Kim, Yeongheon Mok, and Joonki Paik. "Real-Time Surveillance System for Analyzing Abnormal Behavior of Pedestrians". In: *Applied Sciences* 11.13 (July 2, 2021), pp. 6153–. DOI: 10.3390/app11136153.

[21] Feng-Ping An, Jun-e Liu, and Lei Bai. "Pedestrian Reidentification Algorithm Based on Deconvolution Network Feature Extraction-Multilayer Attention Mechanism Convolutional Neural Network". In: *Journal of Sensors* 2021.1 (Jan. 28, 2021), pp. 1–12. DOI: 10.1155/2021/9463092.

[22] Jung-Chul Park, Hyuk Jae Kwon, and Chul Woon Chung. "Innovative digital tools for new trends in teaching and assessment methods in medical and dental education". In: *Journal of educational evaluation for health professions* 18 (June 29, 2021), pp. 13–13. DOI: 10.3352/jeehp.2021.18.13.

[23] Olivier Oheka and Chunling Tu. "Fast and Improved Real-Time Vehicle Anti-Tracking System". In: *Applied Sciences* 10.17 (Aug. 27, 2020), pp. 5928–. DOI: 10.3390/app10175928.

[24] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. "Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods". In: *Journal of Artificial Intelligence Research* 71 (Aug. 30, 2021), pp. 1183–1317. DOI: 10.1613/jair.1.11688.

[25] Xiaowei Jiang, Yuwei Du, and Yingying Zheng. "Evaluation of physical education teaching effect using Random Forest model under artificial intelligence." In: *Heliyon* 10.1 (Dec. 10, 2023), e23576–e23576. DOI: 10.1016/j.heliyon.2023.e23576.

[26] Jayaprada A.B.C., K P Ihalapathirana, Anjalie Gamage, Thilini Jayalath, U L Jayasundara, and Trishan Piyumal. "Player Fitness and Team Squad Selection using Intelligence System in Sri Lanka". In: *International Journal of Computer Applications* 183.36 (Nov. 20, 2021), pp. 54–60. DOI: 10.5120/ijca2021921755.

[27] Chen Xuejuan, Ji-Xiang Du, and Hong-Bo Zhang. "Lipreading with DenseNet and resBi-LSTM". In: *Signal, Image and Video Processing* 14.5 (Jan. 24, 2020), pp. 981–989. DOI: 10.1007/s11760-019-01630-1.

[28] Yu Tan, Wei Su, Lijun Zhao, Qinghui Lai, Chenglin Wang, Jin Jiang, Yongjie Wang, and Peihang Li. "Navigation path extraction for inter-row robots in Panax notoginseng shade house based on Im-YOLOv5s." In: *Frontiers in plant science* 14 (Oct. 17, 2023), pp. 1246717–. DOI: 10.3389/fpls.2023.1246717.

[29] Daniel Maxwell, Sarah L. Meyer, and Charlotte Bolch. "DataStory: an interactive sequential art approach for data science and artificial intelligence learning experiences". In: *Innovation and Education* 3.1 (Oct. 18, 2021), pp. 1–13. DOI: 10.1186/s42862-021-00015-x.

[30] Ehtesham Hassan. "Exploiting multimedia content: a machine learning based approach". In: *ELCVIA Electronic Letters on Computer Vision and Image Analysis* 13.2 (June 9, 2014), pp. 69–69. DOI: 10.5565/rev/elcvia.598.

[31] Qaisar Abbas, Mostafa E. A. Ibrahim, and M. Arfan Jaffar. "Video scene analysis: an overview and challenges on deep learning algorithms". In: *Multimedia Tools and Applications* 77.16 (Dec. 9, 2017), pp. 20415–20453. DOI: 10.1007/s11042-017-5438-7.

[32] R. Rizal Isnanto, Adian Fatchur Rochim, Dania Eridani, and Guntur Cahyono. "Multi-Object Face Recognition Using Local Binary Pattern Histogram and Haar Cascade Classifier on Low-Resolution Images". In: *International Journal of Engineering and Technology Innovation* 11.1 (Jan. 20, 2021), pp. 45–58. DOI: 10.46604/ijeti.2021.6174.

[33] Luyu Xia. "Research on AI Commercial Application development". In: *Highlights in Business, Economics and Management* 24 (Jan. 22, 2024), pp. 1732–1738. DOI: 10.54097/2dnzpg36.

[34] Weiyun Zhao. *Towards a framework for recognising Sign language alphabets captured under arbitrary illumination.* Mar. 18, 2018. DOI: 10.7287/peerj.preprints.26725v1.

[35] 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.* 2024, pp. 7559–7568.

[36] Sudheer Kumar T and Naga Venkateshwara Rao Kollipara. "Real-Time Scene Text Detection Using Deblurring Technique". In: *International Journal of Engineering Technology and Management Sciences* 4.5 (Sept. 28, 2020), pp. 14–21. DOI: 10.46647/ijetms.2020.v04i05.004.

[37] Susmita Goswami et.al. "A Survey on Human Detection using Reinforcement Learning". In: *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.6 (Apr. 11, 2021), pp. 123–126. DOI: 10.17762/turcomat.v12i6.1276.

[38] Marcin Kopaczka, Lukas Breuer, Justus Schock, and Dorit Merhof. "A Modular System for Detection, Tracking and Analysis of Human Faces in Thermal Infrared Recordings." In: *Sensors (Basel, Switzerland)* 19.19 (Sept. 24, 2019), pp. 4135–. DOI: 10.3390/s19194135.

[39] Salah Uddin Yusuf and Azmol Ahmed Fuad. "Real Time Implementation of Face Recognition based Smart Attendance System". In: *WSEAS TRANSACTIONS ON SIGNAL PROCESSING* 17 (July 5, 2021), pp. 46–56. DOI: 10.37394/232014.2021.17.6.