CLASSICALLIBRARY : Pages:1–19

Original Research

Hierarchical Mask Composition for High-Resolution Text-Guided Image Editing

Sanjay Kumar Adhikari[1] and Prerana Shrestha[2]

[1]Department of Information Technology, Far Western University, Mahendranagar–Bhasi Road, Kanchanpur 10400, Nepal.
[2]Department of Computer Applications, Madan Bhandari Memorial College, Bhaktapur–Tokha Road, Kathmandu 44600, Nepal.

Abstract

Text-guided image editing has become central to interactive visual content creation, as natural language offers a flexible interface for specifying semantic modifications. High-resolution editing, however, remains challenging because edits must remain spatially coherent, respect object boundaries, and preserve global structure while responding to localized textual instructions. Existing approaches often rely on a single mask or uniform conditioning over the image, which can lead to spatial bleeding of edits, loss of fine-scale detail, or inconsistent behavior across resolutions. This work introduces a hierarchical mask composition framework for text-guided image editing that decomposes the image plane into a tree of overlapping and nested regions, each associated with distinct textual attributes, editing strengths, or diffusion schedules. The framework constructs a hierarchy of masks ranging from coarse semantic partitions down to fine-grained structures, and composes them in a consistent way to control how local edits propagate across scales. By coupling this hierarchical representation with text-conditioned generative models, the approach enables localized edits at high resolution while maintaining compatibility with latent-space diffusion backbones. The study analyzes the algebraic properties of the composition operator, the numerical behavior of gradient-based optimization of soft masks, and the interaction between hierarchical masking and multi-scale feature representations. Empirical observations on diverse editing tasks indicate that hierarchical mask composition can provide finer spatial control, improved boundary fidelity, and more predictable edit locality compared to single-layer masking strategies, particularly when images are edited at substantially higher resolutions than those used during model pretraining.

1. Introduction

Text-guided image editing connects natural language with visual content, allowing users to specify semantic transformations such as changing object appearance, inserting or removing elements, or modifying global style. Modern generative models, particularly latent diffusion architectures, offer mechanisms to condition the generative trajectory on text embeddings, making it possible to steer image synthesis and editing with relatively high fidelity to a textual prompt. Despite the rapid progress in generative modeling, precise spatial control remains an issue, especially when editing high-resolution images that feature fine structures, complex occlusions, and multiple interacting objects. A typical workflow relies on a mask that indicates the region of interest, which is then combined with the original image and the edited output to produce a final result. While conceptually simple, a single-layer mask is often insufficient to capture the hierarchical organization of visual scenes and the corresponding hierarchy of textual instructions.

High-resolution editing accentuates the limitations of flat masking [1]. At larger resolutions, the semantic content of an image tends to exhibit nested structures, where objects contain parts, parts contain subparts, and textures exhibit multi-scale statistics. Text prompts often reflect this hierarchy implicitly, for example by describing an object and then specifying properties of its parts, or by imposing global stylistic constraints that should apply everywhere except in certain protected regions. A flat mask is forced to approximate this structure with a single binary or soft map, which cannot simultaneously encode priority

rules, partial overlaps, or scale-dependent influence of different text conditions. As a result, edits may leak outside intended regions, global consistency may be disrupted by local modifications, and nuanced instructions may be lost.

Hierarchical mask composition addresses these issues by explicitly modeling the image domain as a hierarchy of regions arranged in a tree or more general directed acyclic structure. Each node in the hierarchy corresponds to a mask defined over the spatial domain of the image and is associated with a textual condition, an editing schedule, or a strength parameter that quantifies how strongly the generative model should respond to that node. Parent nodes represent coarse regions or global constraints, while child nodes represent more localized modifications, refinements, or exceptions. The core question is how to compose these masks into a final effective conditioning signal that modulates the generative process at different spatial locations and scales without introducing inconsistencies.

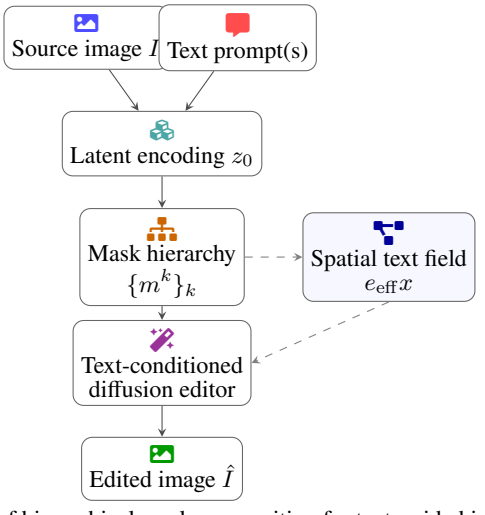

Figure 1: High-level overview of hierarchical mask composition for text-guided image editing. The source image and textual instructions are encoded into a latent representation, which is modulated by a hierarchy of spatial masks. The hierarchy induces an effective, spatially varying text-conditioning field that steers a diffusion-based editor to produce the final edited image while maintaining control over where and how edits are applied.

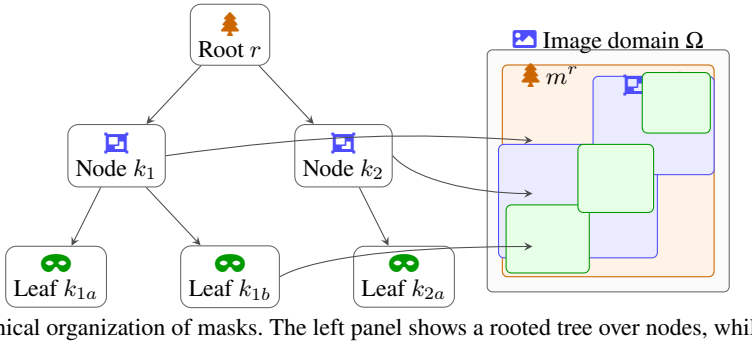

Figure 2: Hierarchical organization of masks. The left panel shows a rooted tree over nodes, while the right panel illustrates overlapping mask fields on the image domain derived from the same hierarchy. Coarse nodes define large regions with broad semantic influence, and deeper nodes refine these regions with more localized masks, enabling structured control over where textual attributes are applied.

Aspect	Description	Role in framework
Spatial control	Text-guided edits localized in space	Prevent leakage of edits outside target regions
High resolution	Edits on large $H \times W$ images	Preserve fine structures and global layout
Hierarchy	Tree of overlapping regions	Encode object/part/scene organization
Mask composition	Operator $\mathcal{C}$ on $\{m^k\}$	Resolve overlaps and precedence across nodes
Latent diffusion	Text-conditioned generative backbone	Realize edits in a multi-scale latent space

Table 1: Key ingredients of the hierarchical mask composition approach for text-guided image editing.

Level	Example region	Typical textual instruction
Root	Whole image	"Render the scene in watercolor style."
Object	Person, building, tree	"Make the dress red and glossy."
Part	Face, window, foliage	"Add freckles to the face."
Sub-part	Eye, pane, leaf cluster	"Brighten the left eye only."
Boundary band	Narrow transition zones	"Blend foreground and background smoothly."

Table 2: Illustrative hierarchy levels and associated text instructions.

Scheme	Effective mask	Main effect
Multiplicative	$\tilde{m}^k = m^k \prod_{j \in \pi k \setminus \{k\}} a^j$	Strong child override of parents
Normalized	$\tilde{m}^k = \frac{s^k m^k}{\sum_j s^j m^j + \varepsilon}$	Bounded total conditioning per pixel
Top-down	Residual mass pushed along tree	Conserves influence from coarse to fine
Depth-weighted	$s^k = \gamma^{h_k}$	Adjusts precedence by node depth
Hybrid	Combination per node or level	Tailors behavior to editing scenario

Table 3: Representative composition schemes for converting intrinsic masks into effective influences.

A hierarchical approach is particularly relevant for high-resolution editing because the underlying generative models often operate in a multi-scale manner. Convolutional and attention-based backbones progressively transform features across resolutions, and diffusion schedulers generate structures from coarse noise to fine detail. Introducing a mask hierarchy enables alignment between semantic structure and architectural scale: coarse masks can influence early diffusion steps or low-resolution layers, while fine masks shape later steps or high-resolution layers. The composition operator thus needs to account not only for spatial overlaps but also for the temporal and depth dimensions of the generative process.

This work develops a formal model of hierarchical mask composition, in which masks are treated as spatial fields and their interactions are governed by algebraic rules that encode precedence, blending, and conflict resolution [2]. Soft masks provide differentiability, making it possible to optimize mask parameters jointly with generative model parameters or with latent codes under reconstruction and text alignment losses. The analysis examines the properties of such operators, including symmetry,

Property	Normalized operator	Implication
Nonnegativity	$\tilde{M}_{ik} \geq 0$ for $M_{ik} \geq 0$	Masks remain valid influence fields
Bounded sum	$_k\, s_k \tilde{M}_{ik} \leq 1$	Interpretable as convex weighting
Coupled gradients	$\partial \tilde{M}_{i\ell} \partial M_{ik}$ depends on all k	Overlaps couple updates across nodes
Stability	Denominator d_i prevents blow-up	Avoids unstable scaling at dense pixels
Depth control	Choice of s_k and hierarchy depth	Trades global consistency vs locality

Table 4: Analytical properties of the normalized hierarchical composition operator.

Loss	Definition sketch	Encouraged behavior	Evaluation region
ℓ_{txt}	Image-text similarity per node	Strong alignment with prompts	Where $\tilde{m}^k$ is large
ℓ_{id}	Distance to original image	Content preservation	Complement of active masks
$\mathcal{R}_{\text{TV}}$	Total variation on m^k	Spatial smoothness of masks	Full image grid
Area penalty	Sum of mask values	Compact, sparse supports	High-variance regions
Hierarchy penalty	Child $\not\leq$ parent violations	Structural consistency	Along parent-child pairs

Table 5: Typical loss components used when optimizing hierarchical masks and latent variables.

Level	Resolution	Dominant nodes	Editing effect
Low	Coarse feature maps	Root, shallow parents	Global style and layout
Mid	Intermediate maps	Object-level nodes	Shape and material of objects
High	Near image scale	Part and boundary nodes	Fine details and edges
Refinement	Optional upsampling	Narrow transition nodes	Seamless blending of regions

Table 6: Coupling between hierarchy depth and multi-scale feature maps in a diffusion backbone.

associativity under certain conditions, and stability under perturbations. This formulation allows the study of numerical issues that arise when many masks overlap, such as gradient attenuation or saturation, and suggests parameterizations that alleviate these issues.

The following sections present background on text-guided image editing and masking strategies, describe the hierarchical mask composition framework and its integration with text-conditioned generative models, analyze the mathematical structure of the compositional operators, and discuss optimization procedures for learning or refining the hierarchy in a data-driven or user-in-the-loop setting. Experimental observations are then described, focusing on qualitative and quantitative aspects of edit locality, boundary sharpness, and consistency across resolutions. The paper concludes with a discussion of limitations and potential directions for extending hierarchical mask composition to more complex scene representations.

Scenario	Target behavior	Hierarchical strategy	Failure mode
Local color change	Restrict edit to one garment	Strong child mask on garment	Color bleeding at weak boundaries
Object insertion	Add object in fixed region	Leaf node with custom prompt	Misalignment with shadows
Style plus edit	Global style + local override	Root style, child exception nodes	Style dominates local change
Noisy segmentation	Correct rough masks	Child refinements with TV regularization	Residual artifacts in cluttered areas
Upscaled editing	Consistent behavior across scales	Shared hierarchy across resolutions	Tile discontinuities if ignored

Table 7: Editing scenarios and typical behaviors observed with hierarchical mask composition.

Aspect	Advantage	Limitation	Possible extension
Spatial precision	Improved locality of edits	Higher implementation complexity	Learnable hierarchy from data
Boundaries	Better transition handling	Requires careful mask design	Dedicated boundary nodes
Optimization	Joint tuning of masks and latents	Gradient attenuation with overlap	Adaptive strengths and pruning
Scalability	Coarse-to-fine control	Overhead in deep trees	Sparse and tile-aware masks
Generalization	Shared rules across images	Depends on backbone capacity	Integration with 3D or multi-view models

Table 8: Summary of benefits, limitations, and extensions suggested by the hierarchical mask formulation.

2. Background on Text-Guided Image Editing and Masking

Text-guided image editing can be framed as the problem of transforming a source image into a target image that reflects a new textual description while preserving certain aspects of the original content. In many pipelines, the source image is encoded into a latent representation, which is then evolved under the influence of a text-conditioned generative process. Masks often enter this pipeline in at least two ways. First, a mask can specify which regions are allowed to change, while other regions are constrained to remain close to the original image. Second, a mask can be used to spatially modulate conditioning signals, such as cross-attention maps or feature injections that depend on the text. While classical photo editing applications used hand-crafted alpha masks for blending, recent generative approaches treat masks as first-class conditioning signals that interact with the learned model.

Let an image be represented as a tensor [3]

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

where H and W denote spatial dimensions and C is the number of channels. A standard mask is a map

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

where m_{ij} indicates the degree to which the pixel at spatial location i, j should be subject to editing. In the simplest form, m is binary, so the edited image $\hat{I}$ is formed by

$$\hat{I}_{ijc} = m_{ij}\, I^{e}_{ijc} \;\; 1 - m_{ij}\, I^{o}_{ijc},$$

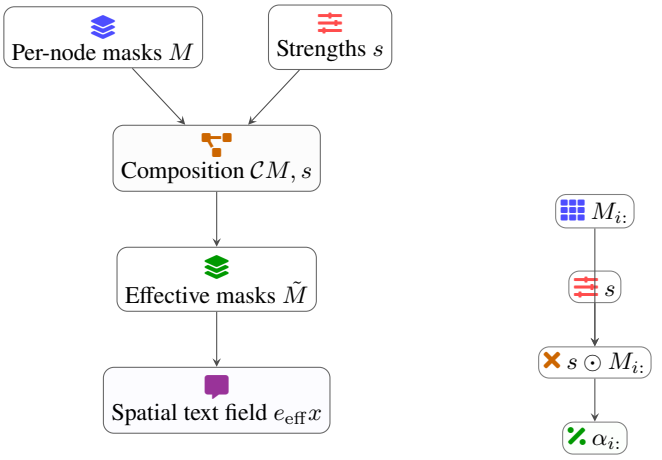

Figure 3: Schematic of the mask composition operator. Per-node masks and scalar strengths are combined by a differentiable operator that enforces bounded total influence at each pixel, yielding effective masks $\tilde{M}$. These effective masks are then used to form a spatially varying text-conditioning field $e_{\text{eff}}x$, which weights textual embeddings according to the hierarchical structure and local mask activations.

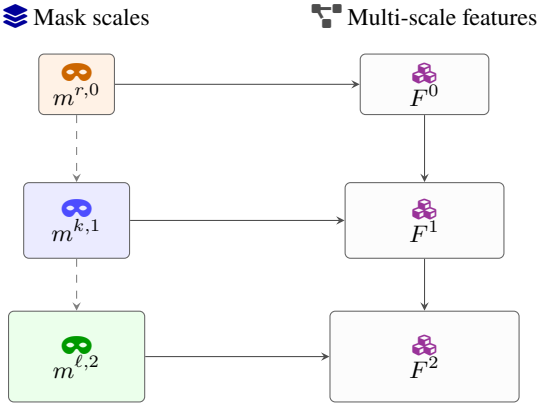

Figure 4: Alignment between hierarchical masks and multi-scale feature representations. Different levels of the mask hierarchy are associated with different resolutions, and each level modulates the corresponding feature maps within a diffusion or U-Net backbone. Coarse masks primarily influence low-resolution features, while finer masks refine high-resolution layers, enabling edits that are coherent across scales yet localized in space.

where I^e is an edited candidate and I^o is the original image. This formulation decouples the synthesis phase from the blending phase and assumes that the generative model can synthesize plausible content in masked regions while leaving the unmasked regions untouched. However, for text-guided editing, the generative model itself can be conditioned on the mask, such that synthesized content responds more strongly inside the mask than outside.

Modern diffusion-based editors commonly encode the image into a latent space

$$z \in \mathbb{R}^{h \times w \times d},$$

where $h < H$, $w < W$, and d is a feature dimension. The mask is downsampled or otherwise mapped to a latent-space mask

$$m^z \in 0, 1^{h \times w}.$$

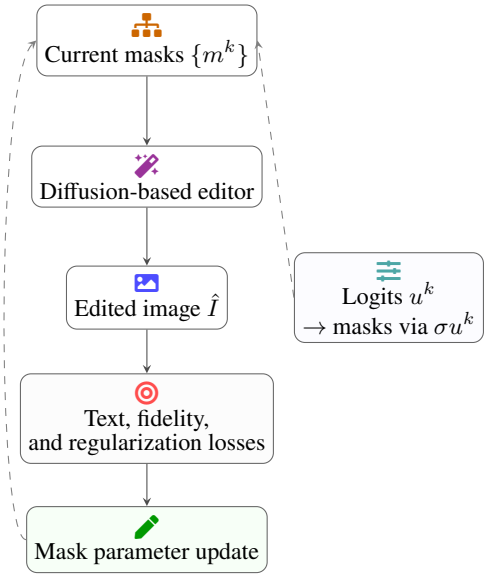

Figure 5: Optimization loop for hierarchical masks. Given an initial hierarchy, a diffusion-based editor produces an edited image whose agreement with text prompts, preservation of protected content, and mask regularity are jointly quantified by a loss. Gradients are backpropagated through the diffusion process and the composition operator to update mask parameters, allowing the hierarchy to adapt its shapes and opacities to better realize the desired edits.

The diffusion model then evolves z under a stochastic differential or difference equation, with drift and noise terms modulated by the textual embedding and possibly by m^z. Because the latent resolution is lower than the image resolution, spatial precision is inherently limited. At the same time, the model usually employs multi-resolution feature maps internally, so information propagates across scales. This creates a mismatch between the single-resolution mask and the multi-scale architecture.

Mask design and manipulation are long-standing topics in image processing and graphics. Classical alpha blending uses a single alpha value per pixel, with simple compositing rules that are associative when colors are interpreted appropriately. Segmentation approaches partition the image into regions that often form hierarchies, where each region may have parent regions representing larger structures [4]. In generative modeling, masks have been used not only for blending but also as latent codes that indicate object presence or absence, or that align with semantic segmentation labels. However, the integration of such hierarchical segmentations into text-guided diffusion editing remains underexplored.

From the perspective of mathematical modeling, a mask is a scalar field over a discrete domain, while the hierarchy of masks can be viewed as a family of scalar fields indexed by a partially ordered set. The composition of masks corresponds to combining these fields according to rules that may depend on their relative positions in the hierarchy, on local values of the fields, and on auxiliary parameters that encode user preferences such as priority or occlusion ordering. The challenge is to define composition operators that are expressive yet sufficiently simple to permit efficient computation and stable optimization.

Existing editing pipelines typically employ flat masks without explicit hierarchy. When multiple masks are used, they are often combined using simple arithmetic operations, for example by taking a minimum, maximum, or weighted sum. These operations do not encode hierarchical precedence beyond simple dominance and do not account for multi-scale architecture. As a result, sudden changes can occur when masks overlap, and the resulting effective conditioning may be difficult to interpret or predict. For high-resolution tasks, these limitations are amplified, since small inconsistencies or artifacts may become more visible.

In contrast, hierarchical structures have proven useful in other areas of vision and graphics, such as image pyramids, wavelet decompositions, and scene graphs. These structures allow algorithms to process

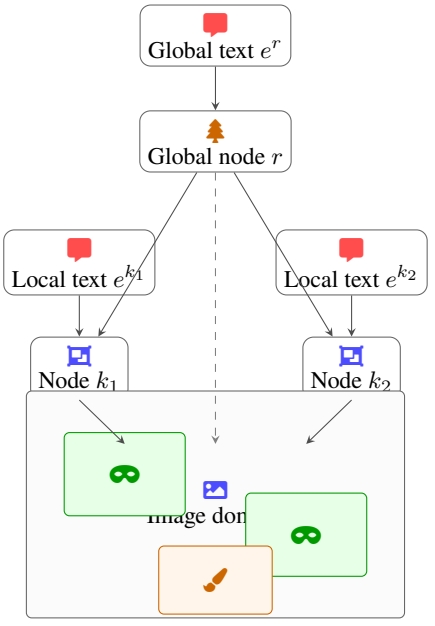

Figure 6: Example of balancing global and local text instructions. A root node carries a global prompt that sets the overall style, while child nodes attach local prompts to specific regions. Hierarchical mask composition allocates conditioning such that the background follows the global instruction, and foreground regions receive stronger influence from their corresponding local prompts, preserving global coherence without sacrificing local edit specificity.

data at different scales, to propagate coarse constraints downwards, and to aggregate fine detail upwards. The goal of hierarchical mask composition for text-guided editing is to bring similar principles to the design of spatial conditioning signals [5]. Rather than treating the mask as an auxiliary artifact detached from the generative architecture, the hierarchy is integrated into both the spatial and scale dimensions of the model, enabling a more faithful translation of multi-level textual instructions into localized visual modifications.

3. Hierarchical Mask Composition Framework

The hierarchical mask composition framework models the image domain as a collection of regions organized in a rooted tree or more general directed acyclic graph. Each node k in this hierarchy is associated with a mask

$$m^k \in 0, 1^{H \times W},$$

a textual embedding e^k, and potentially additional parameters such as editing strength or schedule modifiers. The root node typically corresponds to the entire image or to a global condition, such as a style description that should influence all pixels. Child nodes correspond to subregions that refine the behavior within the parent region, for example specifying edits to objects or parts of objects. The construction of the final effective mask at each pixel position involves combining the contributions of all nodes whose masks assign nonzero values at that position, respecting the hierarchy and parameters that encode precedence.

Formally, consider a finite index set

$$\mathcal{K} = \{1, \ldots, K\}$$

of nodes, with a partial order $\preceq$ that captures the hierarchy. For simplicity, one can assume a rooted tree with root r, although the formulation extends to forests or more general acyclic graphs. Each node

k has a parent $\mathrm{par}k$ except for the root, and a set of children $\mathrm{ch}k$. The mask m^k is interpreted as the intrinsic influence of node k at each pixel, before considering interactions with other nodes. To model precedence, an attenuation factor can be applied to the parent mask in regions where children are active, and conflicting instructions between siblings can be resolved by normalized weighting.

To construct the effective per-node influence at a pixel location x, consider the path from the root to node k. Denote by πk the set of nodes on this path. For a purely multiplicative attenuation model, the effective mask of node k at x can be written as [6]

$$\tilde{m}^k x = m^k x \prod_{j \in \pi k \setminus \{k\}} a^j x,$$

where $a^j x$ is an attenuation factor derived from the children of node j. For example, if the children of j are encouraged to override their parent, then the parent attenuation at location x might be

$$a^j x = 1 - \sum_{\ell \in \mathrm{ch} j} m^\ell x \, w^\ell,$$

where w^ℓ are weights satisfying suitable bounds to ensure nonnegativity. This formulation encodes the idea that where a child is active, its parent influence is reduced. However, pure multiplicative models can suffer from over-attenuation when many levels of the hierarchy overlap.

An alternative is to normalize influences at each pixel so that the total conditioned strength remains bounded. Define unnormalized contributions

$$c^k x = s^k \, m^k x,$$

where s^k is a scalar strength parameter for node k. One can then define normalized effective contributions by

$$\alpha^k x = \frac{c^k x}{\sum_{j \in \mathcal{K}} c^j x \; \varepsilon},$$

where $\varepsilon > 0$ is small. The normalized weights $\alpha^k x$ sum to at most one, depending on the additive structure and ε. To incorporate hierarchy, the scalar strengths s^k can be derived from the depth of the node, from user-specified priorities, or from learned parameters. For example, deeper nodes might be given higher strength to encourage more localized edits, or the opposite if global consistency is prioritized.

A more expressive scheme combines hierarchical attenuation with normalization. At each pixel location, a top-down pass can be performed through the hierarchy, where parent influence is partially allocated to children based on their masks and strengths, and residual influence remains attached to the parent. This can be written as a recursion. For each node k, define a residual influence map

$$r^k x \in 0, 1,$$

initialized at the root as $r^r x = 1$. For each child $\ell \in \mathrm{ch}k$, define an allocation

$$a^{k \to \ell} x = r^k x \, q^\ell x,$$

where $q^\ell x$ is a normalized mask over children, for example

$$q^\ell x = \frac{m^\ell x}{\sum_{j \in \mathrm{ch}k} m^j x \; \varepsilon}.$$

The residual influence for the child is set to $r^\ell x = a^{k \to \ell} x$. The effective mask of node k becomes [7]

$$\tilde{m}^k x = s^k \, r^k x,$$

and the total influence is conserved by the recursive allocation. In this formulation, the hierarchy determines the flow of influence from coarse to fine regions, while masks and strengths control where and how much of that influence is realized.

Once the effective masks $\tilde{m}^k$ are computed, they must be coupled to the text-conditioned generative model. Let the textual embedding associated with node k be

$$e^k \in \mathbb{R}^{d_e},$$

and denote the collection of embeddings by a matrix

$$E \in \mathbb{R}^{K \times d_e}.$$

During diffusion-based editing, cross-attention layers typically compute attention maps between spatial features and text tokens. The hierarchical masks can modulate these interactions. For example, for a spatial location x and a text token associated with node k, the attention logits can be shifted by a factor proportional to $\tilde{m}^k x$, amplifying or suppressing the influence of that textual component. Alternatively, a combined embedding at each spatial location can be formed as a weighted sum

$$e_{\text{eff}} x = \sum_{k \in \mathcal{K}} \tilde{m}^k x \, e^k,$$

which is then passed to the network as a local conditioning vector. This approach effectively maps the hierarchy of masks and textual embeddings to a continuous field of textual features over the image domain.

The hierarchy can also be aligned with architectural scale. Suppose the generative model uses feature maps at resolutions

$$H_0, W_0, H_1, W_1, \ldots, H_L, W_L, 8$$

with $H_0 = H$ and $W_0 = W$. For each level ℓ, a set of masks

$$\tilde{m}^{k,\ell} \in 0, 1^{H_\ell \times W_\ell}$$

can be instantiated by downsampling or learned separately. Coarse nodes in the hierarchy may primarily affect low-resolution levels, while fine nodes influence high-resolution levels. The composition operator is then applied per level, possibly with level-dependent parameters. This allows edits that have global stylistic impact to be injected early in the diffusion process at coarse scales, while localized structural edits are applied later.

From an implementation standpoint, hierarchical mask composition must balance expressiveness with computational overhead. Direct evaluation of recursive allocations for all pixels and nodes may be expensive if the hierarchy is large. However, many practical hierarchies are relatively shallow, and nodes may be confined to particular spatial regions, enabling sparse representations. Furthermore, effective masks can be precomputed once per editing session, rather than being recomputed at every diffusion step, unless the editing schedule requires time-dependent mask modulation.

The framework is flexible with respect to how the hierarchy is obtained. Masks can be provided by users, derived from semantic segmentation models, or initialized from rough scribbles and then refined via optimization. The hierarchy itself can be designed manually or inferred automatically, for example by clustering segments or by performing a recursive partition of the image based on texture and color cues [9]. The composition rules then enforce consistent behavior irrespective of how the hierarchy was created, enabling a uniform interface for text-guided editing across different sources of spatial structure.

4. Mathematical Formulation and Analysis

The hierarchical mask composition framework can be formalized using linear algebra and measure-theoretic language on discrete domains. Consider a discrete spatial domain

$$\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\},$$

and a finite index set $\mathcal{K}$ of nodes. Each mask m^k can be viewed as a vector in

$$\mathbb{R}^{|\Omega|},$$

by flattening spatial indices. The collection of all masks defines a matrix

$$M \in \mathbb{R}^{|\Omega| \times K},$$

whose k-th column is the vectorized mask m^k. Similarly, effective masks $\tilde{m}^k$ form a matrix $\tilde{M}$. The composition operator can then be regarded as a mapping

$$\mathcal{C} : \mathbb{R}^{|\Omega| \times K} \times \Theta \to \mathbb{R}^{|\Omega| \times K},$$

where Θ collects scalar parameters such as strengths and hierarchy structure. The goal is to design $\mathcal{C}$ such that certain desirable properties are satisfied.

One basic property is nonnegativity. For any input M with nonnegative entries, the output $\tilde{M}$ should satisfy $\tilde{M}_{ik} \geq 0$ for all pixels i and nodes k. Another property is boundedness of total influence at each pixel. Let

$$s \in \mathbb{R}^K$$

be a vector of strengths with nonnegative entries. Define the total influence at pixel i as

$$u_i = \sum_{k=1}^{K} s_k \, \tilde{M}_{ik}.$$

It is often desirable to have [10]

$$u_i \leq 1$$

for all i, so that the combined conditioning can be interpreted as a convex combination or as a bounded modulation factor. The normalization-based composition described earlier can be expressed in vector-matrix form as

$$\tilde{M}_{ik} = \frac{s_k \, M_{ik}}{\sum_{j=1}^{K} s_j \, M_{ij} \;\; \varepsilon}.$$

In matrix notation, defining an elementwise product operator $\odot$ and a vector

$$d \in \mathbb{R}^{|\Omega|}$$

with components

$$d_i = \sum_{j=1}^{K} s_j \, M_{ij} \;\; \varepsilon,$$

one can write

$$\tilde{M} = M \, \mathrm{diag} s \oslash d\mathbf{1}^\top,$$

where $\oslash$ denotes elementwise division and $\mathbf{1}$ is a vector of ones of length K. This operator is differentiable wherever $d_i > 0$ and is straightforward to implement.

To encode hierarchy, the strengths s_k can be made dependent on depth or on parent-child relationships. Let h_k be the depth of node k, with $h_r = 0$ at the root. A simple exponential depth weighting

$$s_k = \gamma^{h_k},$$

with $\gamma \in 0, 1$, reduces the effect of deeper nodes when $\gamma < 1$ or emphasizes them when $\gamma > 1$. Alternatively, one can introduce a stochastic interpretation by viewing the normalized influences as probabilities [11]. For each pixel i, the vector

$$p_i = \alpha_{i1}, \ldots, \alpha_{iK}$$

with

$$\alpha_{ik} = \tilde{M}_{ik}$$

can be regarded as a discrete distribution over nodes. This suggests probabilistic formulations in which the node at each pixel is sampled according to p_i and the corresponding text embedding is applied. In practice, deterministic weighted averages are generally used, but the stochastic perspective provides insight into regularization strategies.

The hierarchical allocation scheme can be analyzed using tensor calculus. Let

$$R \in 0, 1^{|\Omega| \times K}$$

denote residual influences, with R_{ir} initialized to one for all pixels and other entries to zero. For each parent-child pair k, ℓ, the allocation from k to ℓ at pixel i is

$$A_{i,k\ell} = R_{ik}\, Q_{i,k\ell},$$

where

$$Q_{i,k\ell} = \frac{M_{i\ell}}{_{j \in \mathrm{ch}k}\, M_{ij} \;\; \varepsilon}$$

if ℓ is a child of k and zero otherwise. The residual influence matrix is updated as

$$R_{i\ell} = {}_{k}\, A_{i,k\ell},$$

propagating influence downwards. This recursive definition can be unfolded into a series of matrix operations, but the resulting expressions depend on the specific topology of the hierarchy.

An important mathematical question is how gradients propagate through the composition operator when masks are optimized jointly with generative parameters. Consider a loss function

$$\mathcal{L}\tilde{M}, \theta$$

that depends on effective masks and model parameters θ [12]. The gradient with respect to M is

$$\frac{\partial \mathcal{L}}{\partial M_{ik}} = {}_{\ell}\, \frac{\partial \mathcal{L}}{\partial \tilde{M}_{i\ell}} \frac{\partial \tilde{M}_{i\ell}}{\partial M_{ik}}.$$

For the normalized composition, the derivative can be written as

$$\frac{\partial \tilde{M}_{i\ell}}{\partial M_{ik}} = \delta_{k\ell}\, \frac{s_k}{d_i} - \frac{s_k s_\ell M_{i\ell}}{d_i^2},$$

where $\delta_{k\ell}$ is the Kronecker delta. This expression shows that gradients couple all nodes at each pixel, because a change in any mask entry modifies the denominator d_i. When many masks overlap, the

denominators may become large, shrinking gradients and potentially slowing optimization. This suggests that alternative parameterizations of M, such as using logits passed through a sigmoid, or using sparse masks, can help maintain gradient magnitudes within a useful range.

The hierarchical allocation scheme introduces additional complexity in gradients due to the recursive dependency of residual influences on masks. The chain rule generates products of attenuation and allocation factors along paths in the hierarchy. In deep hierarchies, this can lead to gradient attenuation similar to vanishing gradients in deep networks. To mitigate this, one can constrain the depth of the hierarchy, introduce residual connections that bypass some levels, or design allocation functions that do not strictly multiply residuals along the entire path. For example, using additive rather than purely multiplicative attenuation for certain levels can preserve gradient magnitudes.

From a discrete mathematics perspective, the hierarchical mask composition can be viewed as operating on a labeled tree over the pixel set. For each pixel, the set of nodes with nonzero mask values defines a finite subset of the tree. Composition resolves this subset into a distribution over nodes according to precedence rules [13]. If one restricts the class of masks to indicator functions of regions that form a partition of the image at each level, then the composition reduces to a consistent assignment of pixels to leaves, possibly with mixtures at boundaries. When masks overlap arbitrarily, the composition acts as a soft resolution of conflicts, which may be interpreted as solving a local discrete optimization problem at each pixel.

The continuous limit, in which the image domain is treated as a compact subset of $\mathbb{R}^2$ and masks as measurable functions, offers another perspective. In this setting, composition yields measurable functions that preserve integrability and boundedness. One can define energy functionals that penalize irregularities in masks, such as total variation or Sobolev norms, and study the variational properties of the composition operator. For instance, if masks are optimized to minimize a loss functional plus a regularization term enforcing spatial smoothness, one can analyze existence and stability of minimizers using tools from calculus of variations. While practical algorithms operate on discrete grids, such continuous analyses can guide the design of numerically stable discretizations.

Finally, the interaction between hierarchical masks and the latent diffusion dynamics can be formulated mathematically. Let

$$z_t$$

denote the latent variable at time t in the diffusion process, evolving according to an update rule

$$z_{t1} = z_t \;\; f z_t, e_{\text{eff}}, t \;\; \sigma_t \, \eta_t,$$

where η_t is random noise and e_{eff} is the effective textual conditioning field derived from the hierarchy. The function f may depend on spatial position, so one can write

$$f z_t, e_{\text{eff}}, t x = g z_t x, e_{\text{eff}} x, t,$$

where g is applied pointwise or via convolutional neighborhoods. The influence of hierarchical masks enters through $e_{\text{eff}} x$, which is a linear combination of embeddings weighted by effective masks. The resulting diffusion dynamics are linear in the masks at the level of conditioning, but nonlinear in terms of the generated images. Analyzing the sensitivity of the final output to variations in masks thus involves understanding how perturbations in conditioning propagate through the diffusion trajectory, a problem that can be explored using linear response approximations or by differentiating through the unrolled diffusion steps [14].

5. Optimization and Numerical Methods

Practical use of hierarchical mask composition in text-guided image editing requires numerically stable and efficient optimization procedures. Masks may be provided by users as rough sketches or bounding

boxes, by automatic segmentation models, or by simple geometric shapes. In many scenarios, these initial masks serve as starting points for an optimization that refines their shapes and opacities to better match textual goals and visual plausibility. The optimization is performed jointly with the editing process, in which latent variables and, potentially, some model parameters are adjusted to fit the desired edit.

Let θ denote the parameters of the generative editor, which may include weights of the diffusion model, adapter layers, or text-image alignment modules. Let z_0 be the latent encoding of the original image, and let $\{z_t\}$ be the sequence produced by the diffusion process under hierarchical conditioning. The final decoded image is

$$\hat{I} = D z_T,$$

where D is the decoder. The objective is to minimize a loss

$$\mathcal{L} = \lambda_{\text{txt}}\, \ell_{\text{txt}} \hat{I}, E \;\; \lambda_{\text{id}}\, \ell_{\text{id}} \hat{I}, I \;\; \lambda_{\text{reg}}\, \mathcal{R} M,$$

where E encodes textual instructions at different nodes, ℓ_{txt} measures text-image alignment, ℓ_{id} measures fidelity to the original image in protected regions, and $\mathcal{R}M$ regularizes masks. The scalars $\lambda_{\text{txt}}, \lambda_{\text{id}}, \lambda_{\text{reg}}$ control the trade-offs among these terms.

The text alignment loss ℓ_{txt} can be defined as a function of similarity between image embeddings and text embeddings. For instance, let

$$\phi I$$

be an image encoder and ψe^k be a text encoder applied to node embeddings. A simple form is

$$\ell_{\text{txt}} \hat{I}, E = -\sum_k \omega_k\, s \phi_k \hat{I}, \psi e^k,$$

where s is a similarity measure and $\phi_k \hat{I}$ extracts features from spatial regions where node k has significant effective mask. The weights ω_k allow varying emphasis across nodes [15]. The identity-preserving loss ℓ_{id} can combine pixel-wise differences, feature distances, and structural similarity measures evaluated on regions where editing is not desired.

Regularization of masks is critical to prevent degenerate solutions, such as masks collapsing to zero or saturating to one everywhere. Spatial smoothness can be enforced via a discrete total variation term

$$\mathcal{R}_{\text{TV}} M = \sum_{k\, i,j} \sqrt{m^k_{i1,j} - m^k_{ij}{}^2 \;\; m^k_{i,j1} - m^k_{ij}{}^2}.$$

This term encourages masks to change gradually across pixels. Additional penalties can enforce sparsity or compact support, for example by adding a term proportional to

$$\sum_{k\, i,j} m^k_{ij},$$

which biases masks towards covering smaller areas. Hierarchical consistency can be encouraged by penalizing violations of the ordering implied by the tree, such as a child being active where the parent is inactive.

Optimization proceeds by gradient-based methods. Masks can be parameterized using logits

$$u^k \in \mathbb{R}^{H \times W}$$

and a sigmoid mapping

$$m^k_{ij} = \sigma u^k_{ij},$$

where σ is the logistic function. This ensures that mask values remain within $0, 1$ without explicit projection. The hierarchy composition operator and the diffusion process are differentiable, at least in an

approximate sense when sampling noise is treated as fixed or reparameterized, enabling backpropagation from the loss to the mask parameters and, if desired, to θ [16]. In practice, the diffusion process may be truncated or approximated to reduce computational cost, and gradients may be accumulated across a reduced set of time steps.

Numerical stability is a central concern when optimizing masks in the presence of normalization-based composition. As observed in the gradient expressions, when denominators d_i become large due to overlapping masks with high strengths, gradients with respect to individual mask entries can become small. To counteract this, several strategies can be employed. One approach is to impose an upper bound on the number of overlapping masks at each pixel, for example by pruning nodes whose masks are negligible in a given region. Another approach is to adapt strengths s_k during optimization, either by learning them or by updating them based on overlap statistics, so that the effective normalization remains in a range that produces meaningful gradients.

Learning the hierarchy structure itself is more challenging than learning mask values. One may start with a flat collection of candidate masks derived from segmentation or attention maps and then cluster them into groups that form parent nodes. A discrete optimization problem arises, where cluster assignments must be determined to maximize an alignment objective. Because direct optimization over combinatorial structures is difficult, a relaxed formulation can be used, where each mask has a soft assignment to multiple parent nodes, encoded by coefficients forming a stochastic matrix. The hierarchy emerges as these coefficients become more peaked during optimization. Techniques from matrix factorization and graph clustering can be adapted to encourage tree-like structures, although enforcing strict tree constraints typically requires discrete postprocessing.

In addition to gradient-based approaches, numerical methods from convex optimization and proximal algorithms can be useful when certain parts of the objective are convex in mask variables [17]. For example, if the regularization term and a subset of the loss are convex in M while the remainder is treated as fixed, one can perform proximal updates on M that solve subproblems with closed-form or efficiently solvable solutions. An alternating minimization scheme can be employed, where masks are updated given fixed generative outputs, and then generative parameters are updated given fixed masks. Although the overall problem is nonconvex, such schemes can help navigate the optimization landscape.

Stochastic optimization plays a role when multiple images and prompts are used to learn generic hierarchical composition parameters. In that setting, mask parameters or composition hyperparameters such as depth-dependent strengths and normalization exponents can be treated as global variables shared across data. Mini-batch stochastic gradient descent or related methods can be applied, with gradients aggregated over different editing tasks. Regularization then serves not only to stabilize optimization but also to encourage generalization across images and prompts.

The numerical implementation must also account for memory and time constraints. Hierarchical masks increase memory usage because multiple mask fields at different levels and scales must be stored. Techniques such as mixed precision, sparse storage for masks with limited support, and recomputation of certain intermediate quantities can reduce memory footprint. Parallelization across pixels, nodes, and diffusion time steps is natural and maps well to modern hardware. Exploiting spatial locality in masks, for example by grouping pixels into tiles and processing tiles independently where possible, can further improve efficiency.

6. Experiments and Analysis

The behavior of hierarchical mask composition for high-resolution text-guided image editing can be examined through a set of editing scenarios that probe different aspects of spatial control, boundary handling, and interaction between local and global textual instructions [18]. While comprehensive numeric benchmarking involves large-scale experiments, qualitative and conceptual analysis already highlights characteristic effects that distinguish hierarchical composition from flat masking.

One class of scenarios involves local attribute modification, such as changing the color or texture of a specific object part while maintaining the rest of the image unchanged. For example, consider an image

containing a person wearing clothing with multiple regions, and a prompt that specifies alterations to only one garment. A flat mask that roughly covers the garment may either leave unwanted gaps or extend into neighboring regions. With hierarchical masks, one can define a parent region corresponding to the general area of the person and a child region corresponding to the specific garment. The parent node encodes broader stylistic constraints, while the child node encodes the detailed textual instruction. During editing, the framework allocates influence from the parent to the child in overlapping areas, allowing the garment edit to take precedence while maintaining global consistency where the parent is active and the child is not.

Boundary fidelity is another aspect where hierarchical composition exhibits distinct behavior. In many images, fine structures such as hair, foliage, or intricate textures lie near regions that users may wish to edit. Hard binary masks often introduce visible seams where edited and unedited regions meet, especially at high resolution. Soft masks alleviate this but can lead to ambiguous regions where both original and edited content contribute. In the hierarchical framework, an additional narrow band of intermediate nodes can be introduced along boundaries, with masks that smoothly transition from one region to another. These intermediate nodes can carry text conditions or editing strengths that interpolate between neighboring regions, allowing the generative model to synthesize content that naturally bridges the transition [19]. The composition rules treat these nodes as part of the hierarchy, so they receive influence from both sides while moderating the final output.

A further set of experiments examines the interaction between global and local instructions. For instance, a prompt may specify a global style, such as a particular artistic rendering, and a local edit, such as inserting an object in one corner. Without hierarchical composition, applying the global style uniformly may inadvertently alter the local insert in unintended ways or cause local edits to propagate. With a hierarchical tree, a root node can reflect the global style, while child nodes define regions that are exempt from certain aspects of the style or that receive a modified version of it. By adjusting strengths and allocation parameters, the editor can maintain the global style in the background while preserving the identity of the inserted object, or vice versa. Observing how the generated images change as these parameters are varied provides insight into the expressive capacity of the hierarchy.

Resolution scaling experiments shed light on how hierarchical composition affects consistency when editing images at resolutions substantially higher than those used for training the generative model. In many cases, editing a high-resolution image requires tiling, downsampling, or multi-pass processing. Flat masks may lead to discontinuities across tiles or to inconsistencies when details are synthesized at different scales. In the hierarchical framework, coarse masks at low resolutions guide the overall distribution of edits, while finer masks at high resolutions refine details. By downsampling masks in a controlled manner and reusing the same hierarchy across scales, one can maintain a consistent editing policy throughout the processing pipeline. Visual inspection of results under different scaling factors can reveal whether the hierarchy helps preserve semantics and spatial alignment as resolution changes [20].

Another dimension of analysis concerns robustness to mask inaccuracies. In realistic use cases, masks obtained from automated segmentation or manual sketching may be imprecise. A flat mask that is slightly misaligned can lead to edits affecting unintended areas. Hierarchical composition allows partial correction of such errors by assigning moderate-strength child nodes that refine the mask in problematic regions. Through optimization, these child masks can shrink or expand to better align with the true object boundaries, while the parent masks retain broader coverage. The extent to which the hierarchy can compensate for initial inaccuracies depends on the regularization strength and the richness of the textual cues. Empirical evaluation involves introducing controlled perturbations to initial masks and observing the degree of recovery achieved by the hierarchical optimization.

Quantitative measures, while dependent on specific implementations and datasets, generally fall into categories such as text-image alignment, fidelity to original content, and localization of edits. Text-image alignment can be assessed by encoding the edited images and textual prompts into a joint embedding space and measuring similarity. Fidelity can be measured by comparing edited and original images in regions that are intended to remain unchanged, using metrics that capture perceptual similarity. Localization can be evaluated by measuring how far edited pixels extend beyond intended regions, for example by computing the overlap between difference maps and masks. Hierarchical composition is expected, in

many settings, to yield lower leakage of edits outside target regions compared to flat masking, because it explicitly models precedence and allocates influence accordingly.

An additional aspect is interpretability [21]. Because the hierarchy organizes masks and textual conditions in a structured manner, it becomes feasible to visualize the contribution of each node to the final edit. For each node, one can render a map of effective mask values and a corresponding visualization of how the edit would look if only that node were active. Comparing these visualizations across nodes helps users understand how their instructions are being applied. This contrasts with flat masks, where the mapping from text to spatial effects can be more opaque, especially when multiple prompts are combined. In interactive settings, users can adjust node strengths or modify masks directly and immediately see localized effects, facilitating a more controllable editing process.

Finally, the experiments highlight limitations of the hierarchical approach. Complex scenes with many overlapping objects and subtle lighting effects may require deep hierarchies and numerous nodes to achieve fine-grained control, increasing computational cost and optimization difficulty. When textual instructions are vague or global in nature, the overhead of managing a hierarchy may not provide clear benefits over simpler masking. Additionally, if the underlying generative model has limited capacity to respect spatial conditioning, no mask hierarchy can fully enforce desired behavior. These observations underline that hierarchical mask composition is most beneficial when there is a clear correspondence between textual instructions and spatial structure, and when high-resolution detail and precise localization are central to the editing task.

7. Conclusion

Hierarchical mask composition offers a structured way to control spatial conditioning in high-resolution text-guided image editing. By organizing masks into a hierarchy that reflects the nested and overlapping structure of visual scenes and textual instructions, the framework enables nuanced regulation of where and how edits are applied. The formulation treats masks as scalar fields organized in a tree-like structure and defines composition operators that map intrinsic mask values to effective influences, respecting precedence and conserving total conditioning strength [22]. When integrated with text-conditioned diffusion models and multi-scale architectures, hierarchical masks can align the semantic hierarchy implied by the prompt with the spatial and scale hierarchy inherent in the generative process.

The mathematical analysis describes how the composition operator can be expressed in linear algebraic form, how normalization and hierarchical allocation interact, and how gradients propagate when masks are optimized jointly with generative parameters. Considerations about nonnegativity, boundedness, and stability guide the design of composition rules that remain numerically tractable even when many masks overlap. The use of differentiable parameterizations for masks, together with regularization terms that enforce smoothness and sparsity, allows optimization methods from continuous optimization and numerical analysis to be applied to what is fundamentally a spatially structured editing problem.

From a practical perspective, hierarchical mask composition provides mechanisms for balancing global and local edits, improving boundary handling, and achieving better edit localization at high resolution than is typically possible with flat masks. The framework accommodates user-provided masks, automated segmentations, and learned refinements, and supports interactive workflows where users can adjust node strengths and structures. At the same time, it introduces additional complexity in managing the hierarchy and associated parameters, which must be justified by gains in control and quality for the editing tasks at hand.

Several directions for further work arise from this formulation. One avenue involves learning hierarchy structures directly from data, using priors on scene organization and textual descriptions to infer suitable trees or graphs over regions. Another direction concerns deeper integration with the internals of generative models, for example by associating nodes with layers or channels rather than only with spatial masks, or by using hierarchical masks to modulate not just conditioning signals but also noise schedules and step sizes in the diffusion process. Extending the framework to three-dimensional or multi-view settings, where masks become volumetric fields or are defined on surfaces, is also a natural progression for

applications involving 3D-aware editing. hierarchical mask composition supplies a principled approach to structuring spatial control in text-guided image editing. It blends concepts from image processing, linear algebra, and optimization with the capabilities of modern generative models, yielding a flexible and mathematically grounded tool for handling complex, high-resolution editing tasks where multi-level textual instructions and spatial precision are central considerations [23].

References

[1] J. Wenk, I. Voigt, H. Inojosa, H. Schlieter, and T. Ziemssen, "Building digital patient pathways for the management and treatment of multiple sclerosis.," *Frontiers in immunology*, vol. 15, pp. 1356436–, 2 2024.

[2] D. R. Karger, R. C. Miller, and M. S. Bernstein, "Crowd-powered systems," *KI - Künstliche Intelligenz*, vol. 27, pp. 69–73, 12 2012.

[3] A. Revanur, N. I. Kolkin, D. Agarwal, S. Agrawal, H. Zhang, M. Harikumar, and E. Shechtman, "Image relighting using machine learning," Oct. 2 2025. US Patent App. 18/949,023.

[4] A. Azarfar, Y. Zhang, A. Alishbayli, S. Miceli, L.-J. Kepser, D. van der Wielen, M. van de Moosdijk, J. R. Homberg, D. Schubert, R. Proville, and T. Celikel, "An open-source high-speed infrared videography database to study the principles of active sensing in freely navigating rodents.," *GigaScience*, vol. 7, pp. 1–6, 12 2018.

[5] M. Bohra, "Implementation of garbage litter detection using image processing with novel perspective of software development," *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, pp. 405–413, 4 2021.

[6] S. Xia, L. Gao, Y.-K. Lai, M. Yuan, and J. Chai, "A survey on human performance capture and animation," *Journal of Computer Science and Technology*, vol. 32, pp. 536–554, 5 2017.

[7] C. Acornley, "Icmi - using generative adversarial networks to create graphical user interfaces for video games," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 802–806, ACM, 10 2021.

[8] Y. Jin and Z. Chen, "A fast resource allocation algorithm based on reinforcement learning in edge computing networks considering user cooperation," *Electronics*, vol. 12, pp. 1459–1459, 3 2023.

[9] H. Pang and Z. Wang, "Dueling double deep q network strategy in mec for smart internet of vehicles edge computing networks," *Journal of Grid Computing*, vol. 22, 2 2024.

[10] D. Bogaevskiy, S. Ezhov, P. Fedoseev, D. Butusov, D. H. Elkamchouchi, M. S. Alqahtani, M. Abbas, B. O. Soufiene, and D. Kaplun, "Key sets analysis of compiler vector options for h.264 video compression algorithms implemented on the mips simd architecture," 6 2023.

[11] A. McAndrew, "Seeing once is better than hearing a thousand times.," *IALLT Journal of Language Learning Technologies*, vol. 10, pp. 11–18, 1 2019.

[12] S. H. Park, A. Tjolleng, J. Chang, M. Cha, J. Park, and K. Jung, "Detecting and localizing dents on vehicle bodies using region-based convolutional neural network," *Applied Sciences*, vol. 10, pp. 1250–, 2 2020.

[13] Y.-D. Zhang, "Guest editorial introduction to the special section on weakly-supervised deep learning and its applications.," *IEEE open journal of engineering in medicine and biology*, vol. 5, pp. 393–395, 5 2024.

[14] T. Bakhshi and S. Zafar, "Hybrid deep learning techniques for securing bioluminescent interfaces in internet of bio nano things.," *Sensors (Basel, Switzerland)*, vol. 23, pp. 8972–8972, 11 2023.

[15] N. Tovar, S. S.-C. Kwon, and J. Jeong, "Image upscaling with deep machine learning for energy-efficient data communications," *Electronics*, vol. 12, pp. 689–689, 1 2023.

[16] X. Gong and F. Wang, "Classification of tennis video types based on machine learning technology," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–11, 6 2021.

[17] P. P. Ray, "An overview of webassembly for iot: Background, tools, state-of-the-art, challenges, and future directions," *Future Internet*, vol. 15, pp. 275–275, 8 2023.

[18] P. Gerstoft, Y. Hu, M. J. Bianco, C. Patil, A. Alegre, Y. Freund, and F. Grondin, "Audio scene monitoring using redundant ad-hoc microphone array networks," *IEEE Internet of Things Journal*, vol. 9, pp. 1–1, 3 2022.

[19] Ángel Fernández Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri, "Deepfakes: current and future trends," *Artificial Intelligence Review*, vol. 57, 2 2024.

[20] F. Fernández-Martínez, A. Hernández-García, M.-A. Fernandez-Torres, I. González-Díaz, Álvaro García-Faura, and F. D. de María, "Exploiting visual saliency for assessing the impact of car commercials upon viewers," *Multimedia Tools and Applications*, vol. 77, pp. 18903–18933, 11 2017.

[21] Q. A. Z. Jabbar, "Enjoyable e-learning modules: Developing and evaluating," *International Journal of Computer Applications*, vol. 145, pp. 42–47, 7 2016.

[22] E. Cesanek, S. Shivkumar, J. N. Ingram, and D. M. Wolpert, "Ouvrai opens access to remote virtual reality studies of human behavioural neuroscience.," *Nature human behaviour*, vol. 8, pp. 1209–1224, 4 2024.

[23] W. Shafik, S. M. Matinkhah, and F. Shokoor, "Recommendation system comparative analysis: Internet of things aided networks," *EAI Endorsed Transactions on Internet of Things*, vol. 8, pp. e5–e5, 5 2022.