Robustness of Mask-Conditioned Image Editors Under Ambiguous or Adversarial Prompts

Suman Thapa[1] **and Kiran Shahi**[2]
[1]Department of Computer Science, Pokhara University, Dhungepatan Road, Pokhara 33700, Nepal
[2]Department of Information Technology, Nepal Open University, Manbhawan Road, Lalitpur 44700, Nepal

ABSTRACT Mask-conditioned image editors based on large generative models are widely used to manipulate local image regions while preserving unedited content. In these systems, a user provides an input image, a binary mask, and a natural language prompt that describes the intended modification inside the masked region. Despite their practical relevance, the behavior of such editors under ambiguous or adversarial prompts has not been characterized in a systematic way. Ambiguity naturally arises when language descriptions are under-specified, conflicting, or stylistically overloaded, while adversarial prompts are crafted to cause mislocalization of edits, content leakage beyond the mask, or unexpected semantic transformations. Understanding the robustness of masked editors to these classes of prompts is important for assessing reliability, safety, and predictability in interactive workflows. This paper examines robustness properties of generic mask-conditioned image editors by formulating a mathematical model of the editing operator, defining explicit robustness metrics over masked and unmasked regions, and analyzing the response to both ambiguous and adversarial prompts. The study considers discrete masks, continuous-valued mask relaxations, and different ways of injecting prompt information into the generative pipeline. A probabilistic framework is introduced to capture distributions over prompts, masks, and images, allowing robustness to be treated as an expected or worst-case risk. Several mitigation strategies are discussed, including regularization of cross-mask interactions, robust optimization over prompt neighborhoods, and training objectives that penalize unintended changes outside the mask. The analysis highlights characteristic failure modes and trade-offs between edit strength, semantic controllability, and spatial robustness without assuming any specific model architecture.

KEYWORDS

1. Introduction

Mask-conditioned image editors have become a standard interface for local image manipulation [1]. In their most common form, a user supplies an image, a binary mask that selects a spatial region of interest, and a natural language prompt describing the desired content inside that region. A generative model then synthesizes a new image that should respect three goals. First, the content inside the mask should align with the prompt at a level consistent with the model's general capabilities. Second, the content outside the mask should remain as close as possible to the input image, up to minimal changes required for consistency. Third, global image properties such as color balance, lighting, and geometry should remain coherent after editing. These requirements define a constrained conditional generation problem that combines elements of inpainting, style transfer, and text-guided synthesis.

The success of such editors relies on strong priors encoded by large-scale generative models, including diffusion-based architectures and other latent-variable models. The mask conditions the spatial support of the edit, while the prompt conditions the

Article info:
Suman Thapa and Kiran Shahi. *Robustness of Mask-Conditioned Image Editors Under Ambiguous or Adversarial Prompts.* Grovesocieties .

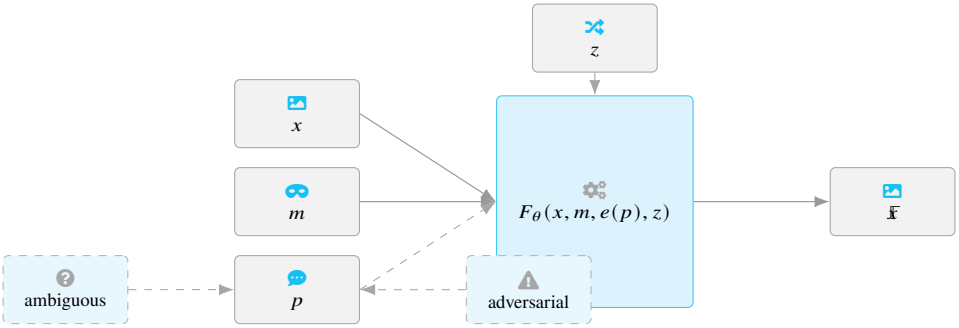

Figure 1 Mask-conditioned editor as a stochastic operator acting on image x, mask m, prompt p, and noise z. Ambiguous and adversarial prompts both act through the language pathway, shaping $e(p)$ and influencing the edited image $\mathbb{F}$ whose localization and semantics are analyzed.

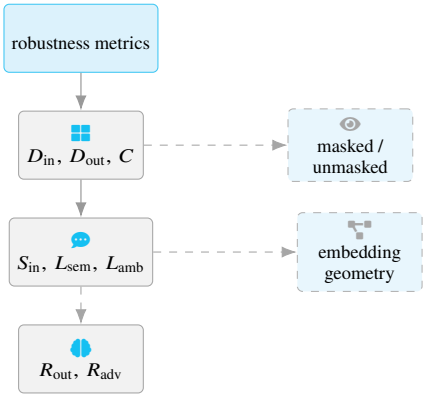

Figure 2 Organization of robustness metrics. Spatial quantities D_{in}, D_{out}, and contrast C focus on localization, semantic quantities S_{in}, L_{sem}, and L_{amb} focus on text–image alignment, and prompt-level risks R_{out}, R_{adv} summarize expected and worst-case degradation as prompts vary.

semantic content. The interaction between these two conditioning channels is not explicitly constrained by most training pipelines, which often rely on generic text-to-image training data together with finetuning or architectural modifications to handle masked editing [2]. As a result, the effective behavior of the editing operator can deviate from user expectations when the prompt is ambiguous, unusual, or deliberately adversarial [3].

Ambiguous prompts appear in several forms. A single phrase may admit multiple plausible visual interpretations due to polysemy, compositional ambiguity, or under-specified attributes such as style and viewpoint. The prompt may also be locally clear but globally ambiguous when combined with the existing image, for example when the requested object conflicts with the scene context. Moreover, natural language descriptions can be rich in artistic or metaphorical terms whose mapping to concrete visual attributes is not uniquely defined. When a mask-conditioned editor encounters such prompts, it must implicitly select one realization among many, and this choice can interact in intricate ways with the mask geometry and with internal attention mechanisms that distribute the prompt's influence across the image.

Adversarial prompts are crafted with the explicit intent of inducing undesired behavior. In the context of mask-conditioned editing, adversarial objectives include spreading the edit beyond the mask, causing drastic background changes, inserting semantically unrelated content, or bypassing safety filters by using obfuscated or indirect phrasing. These adversarial inputs can operate at different abstraction levels, from low-level perturbations in the text embedding space to higher-level manipulations of sentence structure and semantics. Unlike adversarial perturbations on pixel inputs, prompt-based adversaries act through the language pathway and propagate through cross-modal attention and conditioning modules that may not have been exposed to similar inputs during training [4].

The robustness of mask-conditioned editors under such challenging prompts can be studied from both empirical and theoretical perspectives. Empirically, one can measure how often and to what extent edits leak outside the mask, how strongly the background is affected, or how frequently the generated content deviates from the intended semantics. Theoretically, one can attempt to formalize the editing operator as a conditional transformation of an image tensor, with the mask and prompt acting as control variables, and then analyze its stability under perturbations in the prompt space. This perspective connects naturally to concepts from multivariate calculus, probability theory, and robust optimization.

This paper develops a framework for reasoning about robustness in mask-conditioned image editing under ambiguous and adversarial prompts. The approach starts from a mathematical formalization of the editor as a stochastic operator acting on images, masks, and prompt embeddings. Robustness metrics are introduced to quantify localization fidelity, semantic alignment, and invariance of the unmasked region. Ambiguous prompts are modeled as giving rise to sets or distributions of plausible

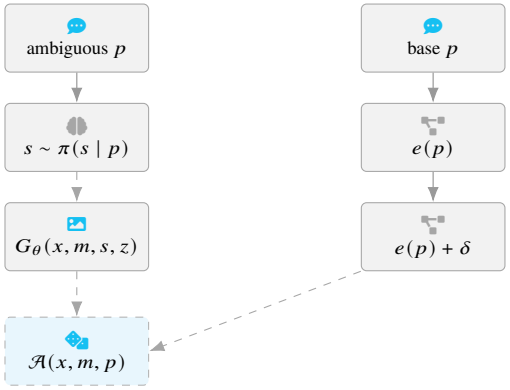

Figure 3 Ambiguous prompts are modeled via latent semantics $s \sim \pi(s \mid p)$ whose decoded images populate the admissible set $\mathcal{A}(x, m, p)$. Adversarial prompts correspond to constrained perturbations δ of the embedding $e(p)$, which can shift the editor toward harmful regions within or outside $\mathcal{A}(x, m, p)$, affecting localization and semantics.

Symbol	Description	Space / type	Role in formulation
x	Input image tensor	$\mathbb{R}^{H\times W\times C}$	Observed conditioning image before editing
m	Binary spatial mask	$\{\varkappa,\}^{H\times W}$	Selects editable region versus background
p	Text prompt (token sequence)	$\mathcal{P}$	Natural language specification of the intended edit
$e(p)$	Prompt embedding	$\mathbb{R}^d$	Continuous text representation used for conditioning the generator
z	Exogenous noise source	$\mathbb{R}^k$, $z \sim \nu$	Captures stochasticity of the generative process
F_θ	Deterministic generator	$\mathbb{R}^{H\times W\times C}\times\{\varkappa,\}^{H\times W}\times\mathbb{R}^d\times \mathbb{R}^k \to \mathbb{R}^{H\times W\times C}$	Maps $(x, m, e(p), z)$ to a single edited image
f_θ	Stochastic editor	$\mathbb{R}^{H\times W\times C} \times \{\varkappa,\}^{H\times W} \times \mathcal{P} \to \mathcal{P}(\mathbb{R}^{H\times W\times C})$	Induced distribution over edited images given (x, m, p)

Table 1 Core variables and operators defining a generic mask-conditioned image editor.

outputs, while adversarial prompts are modeled via worst-case perturbations in prompt space subject to constraints on semantics or magnitude. Within this framework, one can examine how architectural choices and training objectives influence cross-mask interactions and sensitivity to prompt variations.

The analysis does not assume access to model internals for specific implementations, but instead emphasizes structural properties shared by a broad class of mask-conditioned editors. This includes additive or multiplicative mask injection into latent codes, cross-attention conditioning on prompt embeddings, and diffusion-time conditioning strategies. By focusing on properties that are largely architecture-agnostic, the discussion aims to separate aspects of robustness that stem from generic constraints of text-conditioned generative modeling from those that depend on particular design decisions.

Finally, the paper considers mitigation strategies that may improve robustness without fundamentally changing the model class. These strategies include mask-aware losses that penalize undesired edits outside the masked region, regularization of prompt gradients to reduce sensitivity to adversarial perturbations, and training procedures that expose the model to deliberately ambiguous prompts paired with diverse but controlled target outputs. The treatment remains qualitative and analytical, with an emphasis on identifying trade-offs between robustness, edit strength, and expressive power.

2. Background and Problem Formulation

A mask-conditioned image editor can be idealized as a mapping from an input image, a mask, and a prompt to a distribution over edited images. Let an input image be represented as a real-valued tensor with spatial and channel structure. Denote this image by

$$x \in \mathbb{R}^{H\times W\times C} \tag{1}$$

Quantity	Definition	Intuition / geometric role
V	Full pixel grid $\{,\dots,H\} \times \{,\dots,W\}$	Node set when viewing the image as a lattice graph
$M(m)$	Editable index set $\{(i,j) \in V : m_{ij} = \}$	Pixels whose content is allowed to change during editing
$V \setminus M(m)$	Unmasked index set	Pixels that should remain as close as possible to the input
∂M	Mask boundary: masked pixels adjacent to at least one unmasked pixel	Interface across which information can leak during generation
$\lvert M(m)\rvert$	Mask area (cardinality of $M(m)$)	Controls scale of the edit and effective degrees of freedom
$\lvert\partial M\rvert$	Boundary length in pixels	Correlates with risk of halo artifacts and boundary leakage
$\lvert M(m)\rvert/\lvert V\rvert$	Relative mask area	Fraction of the image under direct prompt control

Table 2 Discrete geometric descriptors of the mask and its interaction with the pixel lattice.

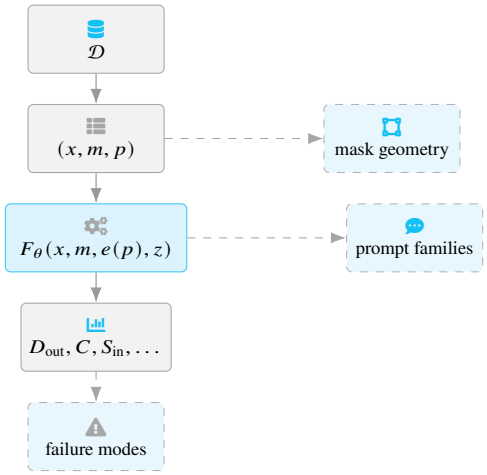

Figure 4 Empirical robustness analysis: samples (x, m, p) from $\mathcal{D}$ are passed through the editor, robustness metrics are computed, and characteristic failure modes are identified. Conditioning on mask geometry statistics and structured prompt families reveals how leakage, global drift, and semantic errors depend on spatial support and linguistic structure.

where the dimensions correspond to height, width, and color channels respectively. The binary mask is modeled as [5]

$$m \in \{, \}^{H \times W} \tag{2}$$

with the convention that pixels with value one belong to the editable region and zeros denote the region that should remain unchanged as much as possible.

The prompt is a natural language sequence drawn from a discrete vocabulary. On the discrete level, it can be represented as a finite sequence

$$p = (t, t, \dots, t_L) \tag{3}$$

where each token belongs to a finite alphabet. For the purposes of conditioning the generative model, this discrete sequence is typically mapped to a continuous embedding. Abstractly, this yields a prompt embedding vector

$$e(p) \in \mathbb{R}^d \tag{4}$$

produced by a text encoder that is often pretrained. The mask-conditioned editor then realizes a stochastic mapping

$$f_\theta : \mathbb{R}^{H \times W \times C} \times \{, \}^{H \times W} \times \mathcal{P} \to \mathcal{P}\big(\mathbb{R}^{H \times W \times C}\big) \tag{5}$$

where θ collects model parameters, $\mathcal{P}$ denotes the space of prompts, and $\mathcal{P}(\cdot)$ the space of probability distributions on images. Given (x, m, p), the editor samples an output image [6]

$$\mathbb{F} \sim f_\theta(x, m, p) \tag{6}$$

through a generative process that may involve iterative denoising, latent space transformations, or other constructions.

To capture the stochasticity more explicitly, it is convenient to write the editor as a deterministic function of the input variables

Category	Example characteristics	Modeling abstraction	Typical effect on the editor
Lexical ambiguity	Polysemous nouns, adjectives with several visual realizations	Multimodal $\pi(s \mid p)$ over latent semantics	Different plausible objects or attributes synthesized across runs
Contextual ambiguity	Prompt conflicts with existing scene or objects	Multiple feasible s conditioned on (x, m, p)	Inconsistent global context or surprising background adjustments
Stylistic ambiguity	Metaphors, artistic or mood-heavy descriptors	Broad prior over style-related dimensions in s	Variability in texture, color palette, or lighting across samples
Adversarial mislocalization	Prompts constructed to spread edits beyond the mask	Optimizer over $p' \in \mathcal{S}(p)$ maximizing leakage	Background edited strongly, reduced localization contrast C
Adversarial safety-bypass	Obfuscated or coded references to disallowed content	Optimization of safety-related loss while evading filters	Harmful semantics synthesized inside or outside the mask
Hybrid ambiguous–adversarial	Ambiguity deliberately exploited in adversarial fashion	Shifts mass within $\mathcal{A}(x, m, p)$ toward bad outputs	Model biased toward undesirable interpretations of the same prompt

Table 3 Qualitative taxonomy of ambiguous and adversarial prompts in mask-conditioned editing.

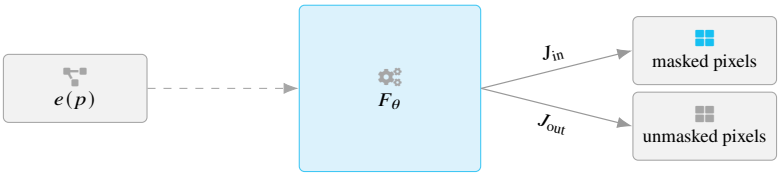

Figure 5 Differential view of prompt sensitivity. The Jacobian with respect to $e(p)$ decomposes into J_{in} affecting masked pixels and J_{out} affecting unmasked pixels. Controlling the effective norm of J_{out} supports local guarantees that small prompt perturbations, including adversarial ones in embedding space, have limited impact on the background.

and an exogenous noise source. Let z be a random vector drawn from a fixed distribution such as a multivariate standard normal, and define a deterministic mapping

$$F_\theta : \mathbb{R}^{H\times W\times C} \times \{\text{ø}, \text{ø}\}^{H\times W} \times \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^{H\times W\times C} \tag{7}$$

such that

$$\text{ø} = F_\theta\big(x,\ m,\ e(p),\ z\big), \qquad z \sim \nu \tag{8}$$

for some fixed noise distribution ν. The mapping F_θ encapsulates the generative procedure, including mask injection, cross-attention operations, and diffusion-time conditioning if a diffusion formulation is used.

The mask determines a discrete subset of spatial locations. Let the pixel grid be the finite set

$$V = \{, \dots, H\} \times \{, \dots, W\} \tag{9}$$

and define the editable set of indices

$$M(m) = \big\{(i, j) \in V : m_{ij} = \big\}. \tag{10}$$

The complement $V \setminus M(m)$ indexes pixels that nominally should not be altered by the editor except for minimal adjustments required by global consistency. This graph-theoretic view of the pixel lattice, with nodes corresponding to pixels and edges connecting spatial neighbors, provides a connection to discrete mathematics [7]. In particular, one can consider the induced subgraph on $M(m)$ and analyze how information flows through the generative network from masked to unmasked nodes via convolutional or attention layers.

For later analysis, it is useful to decompose an image into masked and unmasked components. Define an elementwise product

$$(x \odot m)_{ijc} = x_{ijc} m_{ij} \tag{11}$$

and its complement

$$(x \odot (\ - m))_{ijc} = x_{ijc}(\ - m_{ij}). \tag{12}$$

The edit is intended to primarily affect the component $x \odot m$, while leaving $x \odot (\ - m)$ close to its original values. The edited image ø can be similarly decomposed, and robustness metrics will compare these decompositions.

The data distribution over which the editor operates is modeled as a probability distribution

$$\mathcal{D} \text{ on } \mathbb{R}^{H\times W\times C} \times \{\text{ø}, \text{ø}\}^{H\times W} \times \mathcal{P}. \tag{13}$$

A sample $(x, m, p) \sim \mathcal{D}$ represents a typical user interaction drawn from the deployment scenario. For each triplet, there

Metric	Definition (schematic)	Primary signal	Region focus
D_{out}	$\|(x - F) \odot (\ - m)\|$	Deviation of background from input	Unmasked region preservation
D_{in}	$\|(x - F) \odot m\|$	Magnitude of change within mask	Edit strength inside mask
C	$D_{\text{in}}/(D_{\text{out}} + \varepsilon)$	Ratio of in-mask to out-of-mask change	Spatial localization of edits
S_{in}	Cosine between $g_{\text{img}}(F^{\text{mask}})$ and $g_{\text{txt}}(p)$	Alignment of edited region with prompt semantics	Masked semantic consistency
$R_{\text{out}}(\theta)$	$\mathbb{E}[D_{\text{out}}(x, F; m)]$ over $(x, m, p), z$	Average leakage under deployment prompts	Global background robustness
$R_{\text{adv}}(\theta)$	$\mathbb{E}[\sup_{p' \in \mathcal{S}(p)} D_{\text{out}}(x, F_\theta(x, m, p'), z), m)]$	Worst-case leakage under constrained prompt attacks	Adversarial robustness of background
L_{amb}	$\inf_{u \in U(x,m,p)} L_{\text{sem}}(F, u)$	Distance to best acceptable semantic interpretation	Robustness to ambiguous prompts

Table 4 Spatial and semantic robustness metrics used to characterize mask-conditioned editing behavior.

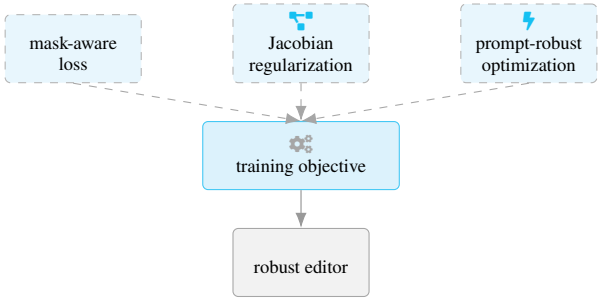

Figure 6 Mitigation strategies combined in the training objective: mask-aware background preservation, Jacobian regularization that constrains prompt-to-background sensitivity, and robust optimization over prompt neighborhoods jointly shape an editor that better balances edit strength, semantic controllability, and spatial robustness.

exists at least one notion of an ideal or target output image, denoted $x^\star$, even if this target is not uniquely defined in practice. Training procedures attempt to adjust θ so that the distribution of $F_\theta(x, m, e(p), z)$ concentrates around plausible targets $x^\star$.

Ambiguity in prompts can be modeled by associating a set of plausible targets with each (x, m, p). Let

$$\mathcal{A}(x, m, p) \subset \mathbb{R}^{H \times W \times C} \tag{14}$$

denote the set of images considered acceptable realizations of the editing request. This set may be defined via semantic constraints, constraints on the masked region, or invariants in the unmasked region. In the presence of ambiguity, $\mathcal{A}(x, m, p)$ may contain many disparate images. When the prompt is precise and the editing task is well specified, $\mathcal{A}(x, m, p)$ may be effectively concentrated.

Adversarial prompts can be thought of as elements of a subset [8]

$$\mathcal{P}_{\text{adv}} \subset \mathcal{P} \tag{15}$$

selected by an adversary with access to the editor. In a more refined view, given a base prompt p, an adversary chooses a perturbed prompt

$$p' \in \mathcal{S}(p) \tag{16}$$

from a constraint set $\mathcal{S}(p)$ that captures syntactic or semantic similarity, such as bounded distance in embedding space or limited token-level modifications. The adversary's goal is to maximize some notion of loss or violation of robustness properties. This adversarial game can be formalized once specific robustness metrics and perturbation sets are defined.

The abstract formulation presented in this section isolates the essential components of a mask-conditioned editor. It highlights the roles of the input image, the discrete mask, and the continuous prompt embedding, as well as the underlying randomness of the generative process. This structure will be used in subsequent sections to define robustness metrics, analyze ambiguous and adversarial prompts, and discuss mitigation strategies based on probabilistic and optimization-theoretic arguments.

3. Robustness Metrics for Mask-Conditioned Editing

Robustness of a mask-conditioned editor concerns how its outputs respond to variations in prompts and noise, while preserving desirable invariances with respect to the unmasked region. To reason about robustness quantitatively, it is necessary to define metrics on the space of outputs that capture spatial localization

Object	Role	Mathematical space	Interpretation
s	Latent semantic code specifying detailed edit	$\mathcal{S} \subset \mathbb{R}^q$	Encodes object identity, style, layout within the masked region
$\pi(s \mid p)$	Conditional distribution over latent semantics	Probability distribution on $\mathcal{S}$	Captures ambiguity induced by the natural-language prompt
$G_\theta(x, m, s, z)$	Decoder conditioned on (x, m, s, z)	$\mathbb{R}^{H\times W\times C}$	Approximate implementation of F_θ with explicit s
$\mu(x, m, p)$	Mean edited image $\mathbb{E}[G_\theta(x, m, s, z) \mid p]$	Image-valued expectation	Average behavior of the editor for an ambiguous prompt
$\Sigma(x, m, p)$	Covariance of $G_\theta(x, m, s, z)$ given p	Linear operator on image space	Per-pixel and cross-pixel variability across interpretations and noise
$\mathcal{A}(x, m, p)$	Acceptable set of edited images	Subset of $\mathbb{R}^{H\times W\times C}$	Ideal outputs satisfying semantic and localization constraints
$U(x, m, p)$	Acceptable semantic embeddings	Subset of $\mathbb{R}^r$	Reference interpretations used in ambiguity-aware losses L_{amb}

Table 5 Latent-semantic and set-based constructs for modeling ambiguity in prompt-conditioned editing.

and semantic behavior. This section introduces several such metrics, expressed in a probabilistic framework [9].

Consider an input triplet (x, m, p) and the corresponding edited image $\mathbb{F}$. A fundamental requirement is preservation of the unmasked region. One way to capture this is through a distance between the unmasked parts of the input and output images. For a norm $\|\cdot\|$ on the space of images, define the unmasked distance

$$D_{\text{out}}(x, \mathbb{F}; m) = \big\|(x - \mathbb{F}) \odot (\ - m)\big\|. \tag{17}$$

The choice of norm can vary. A simple option is the Euclidean norm on the vectorized image, while perceptual norms based on feature embeddings can also be used. For robustness analysis, the precise form is less important than the qualitative requirement that small values of D_{out} correspond to high fidelity outside the mask.

Analogously, the change inside the masked region is measured by

$$D_{\text{in}}(x, \mathbb{F}; m) = \big\|(x - \mathbb{F}) \odot m\big\|. \tag{18}$$

In many editing tasks, one desires sufficiently large changes within the mask, to reflect the requested transformation, together with small changes outside. These competing objectives can be combined into a contrast measure

$$C(x, \mathbb{F}; m) = \frac{D_{\text{in}}(x, \mathbb{F}; m)}{D_{\text{out}}(x, \mathbb{F}; m) + \epsilon} \tag{19}$$

for a small constant $\epsilon > \varkappa$ used to avoid division by zero [10]. Higher values of C indicate that most of the change is concentrated inside the mask. An editor with good spatial robustness will tend to produce large C across a range of prompts.

To incorporate semantics, consider an embedding model that maps images and texts into a joint vector space. Let

$$g_{\text{img}} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^r \tag{20}$$

be an image encoder and

$$g_{\text{txt}} : \mathcal{P} \to \mathbb{R}^r \tag{21}$$

be a text encoder aligned with it. The semantic similarity between an image $\mathbb{F}$ and a prompt p can be measured by the cosine similarity

$$S(\mathbb{F}, p) = \frac{\langle g_{\text{img}}(\mathbb{F}), g_{\text{txt}}(p)\rangle}{\|g_{\text{img}}(\mathbb{F})\|\,\|g_{\text{txt}}(p)\|}. \tag{22}$$

When considering mask-conditioned editing, it is useful to apply the image encoder to masked versions of the image. Denote by x^{mask} the image obtained by zeroing out pixels outside the mask, or by using an inpainting token. Then semantic alignment inside the mask can be estimated as

$$S_{\text{in}}(\mathbb{F}, p; m) = S\big(\mathbb{F}^{\text{mask}}, p\big). \tag{23}$$

Similarly, one can define a semantic drift for the unmasked region by masking the complement of the region and comparing to embeddings associated with the original content.

The robustness of the editor with respect to prompt perturbations can be cast in a probabilistic manner. For a distribution $\mathcal{P}_{x,m}$ of prompts conditioned on (x, m), consider the expected unmasked distance

$$R_{\text{out}}(\theta) = \mathbb{E}\Big[D_{\text{out}}\big(x,\ F_\theta(x, m, e(p), z); m\big)\Big] \tag{24}$$

Failure mode	Description	Typical triggers	Metric signature
Boundary leakage	Edits bleed a few pixels beyond the mask boundary	Large $\lvert\partial M\rvert$, strong local changes, challenging lighting consistency	Moderate D_{out} concentrated near boundary; reduced contrast C
Background drift	Global color, tone, or texture shifts in unmasked region	Style-heavy prompts, large masks, strong global priors	Elevated D_{out} across wide area; semantics roughly preserved
Semantic spillover	Prompt attributes applied to background objects	Complex multi-object prompts, ambiguous modifier scope	Image embeddings of unmasked region correlate with prompt semantics
Mask ignoring	Model replaces or strongly alters entire scene	Adversarial prompts, unusual requests, extreme guidance scales	D_{out} comparable to D_{in}; very low localization contrast C
High output variance	Large run-to-run variability for fixed (x, m, p)	Strong ambiguity in p, high stochasticity, unstable diffusion schedules	High empirical variance of D_{out} and semantic scores over z

Table 6 Characteristic empirical failure modes observed in mask-conditioned image editors and their diagnostic signatures.

where the expectation is taken over $(x, m, p) \sim \mathcal{D}$ and $z \sim \nu$. A low value of $R_{\text{out}}(\theta)$ indicates that on average the editor leaves the unmasked region close to the input, even as the prompt varies according to the deployment distribution. When ambiguity is present, the distribution of prompts may exhibit significant variability [11].

Worst-case robustness to prompt perturbations is captured by an adversarial risk. Given a base prompt p and a constraint set $\mathcal{S}(p)$, define

$$R_{\text{adv}}(\theta) = \mathbb{E}\Big[\sup_{p' \in \mathcal{S}(p)} D_{\text{out}}\big(x,\ F_\theta(x, m, e(p'), z); m\big)\Big]. \quad (25)$$

This quantity measures the largest possible disturbance to the unmasked region achievable by an adversary constrained to stay within $\mathcal{S}(p)$. A small value of $R_{\text{adv}}(\theta)$ reflects robustness against prompt-based attacks that attempt to break localization.

Semantic robustness can be described similarly. For a target semantic embedding u associated with the intended edit, define a semantic loss

$$L_{\text{sem}}(\mathbb{F}, u) = -\frac{\langle g_{\text{img}}(\mathbb{F}^{\text{mask}}), u\rangle}{\|g_{\text{img}}(\mathbb{F}^{\text{mask}})\|\,\|u\|}. \quad (26)$$

Under ambiguous prompts, there may exist multiple acceptable targets u corresponding to different interpretations. For each task, define a set of acceptable semantic embeddings

$$U(x, m, p) \subset \mathbb{R}^r. \quad (27)$$

A semantic robustness measure can then be defined as the distance from the best alignment

$$L_{\text{amb}}(\mathbb{F}; x, m, p) = \inf_{u \in U(x,m,p)} L_{\text{sem}}(\mathbb{F}, u). \quad (28)$$

This quantity is small if the generated image lies close to at least one acceptable semantic interpretation of the ambiguous prompt. A robust editor should keep L_{amb} under control while still respecting spatial constraints.

From a multivariate calculus perspective, robustness can be linked to sensitivity of the output with respect to the prompt embedding. Consider the mapping

$$\Phi_\theta(x, m, e, z) = F_\theta(x, m, e, z) \quad (29)$$

and its Jacobian with respect to the embedding variable [12]

$$J_e(x, m, e, z) = \mathbb{0}_e \Phi_\theta(x, m, e, z). \quad (30)$$

The magnitude of J_e in appropriate norms provides a local measure of how strongly small changes in the embedding affect the image. For robustness against small adversarial perturbations in embedding space, one would like the operator norm of J_e restricted to the unmasked region to be small. Define a linear operator P_{out} that projects an image onto its unmasked components. Then the effective Jacobian is

$$J_{\text{out}}(x, m, e, z) = P_{\text{out}} \circ J_e(x, m, e, z). \quad (31)$$

Bounding the operator norm of J_{out} provides a sufficient condition for local robustness of the background to embedding perturbations.

These metrics combine concepts from linear algebra, probability theory, and multivariate calculus to characterize robustness. They provide a vocabulary for discussing empirical phenomena such as leakage of edits beyond the mask, sensitivity to prompt rephrasing, and the behavior of the editor under ambiguous instructions. In subsequent sections, these measures will be related to models of ambiguous and adversarial prompts and to possible mitigation strategies.

4. Ambiguous and Adversarial Prompt Modeling

Ambiguous prompts arise naturally in interactive image editing workflows. To model ambiguity, it is necessary to distinguish

Strategy	Mechanism	Targeted issues	Trade-offs
Background-preserving loss L_{bg}	Penalize $(x - F_\theta(x, m, e(p), z)) \odot (-m)$ during training	Boundary leakage, global background drift	May under-adjust global lighting or geometry that should change for realism
Jacobian regularization L_{jac}	Constrain $\mathbb{0}_e F_\theta$ on unmasked pixels	Sensitivity to adversarial prompt perturbations	Extra compute; reduced responsiveness to subtle benign prompt changes
Ambiguity-aware objectives	Train against multiple acceptable targets or semantic sets $U(x, m, p)$	Robustness under underspecified or metaphorical prompts	More complex annotation and optimization, potential multimodal outputs
Robust prompt optimization L_{rob}	Inner maximization over $p' \in \mathcal{S}(p)$ during training	Worst-case prompt attacks causing leakage or unsafe behavior	Increased training time; possible over-regularization of edits
Attention regularization	Limit cross-attention mass from prompt to background features	Semantic spillover and mask ignoring	May weaken long-range dependencies and global coherence
Architectural decoupling	Block-structured layers separating masked and unmasked pathways	Structural control of cross-region interactions	Reduced expressiveness for edits requiring coordinated global changes

Table 7 Mitigation techniques aimed at improving robustness of mask-conditioned editors and their qualitative trade-offs.

between uncertainty about the semantic content and uncertainty about stylistic or contextual attributes. A convenient formalism is to represent an ambiguous prompt as inducing a probability distribution over latent semantic variables. Let p be a given ambiguous text description, and let s denote a latent semantic code that fully specifies the content of the edit inside the mask. One can consider a conditional distribution

$$\pi(s \mid p) \text{ on a latent space } \mathcal{S} \subset \mathbb{R}^q. \tag{32}$$

Different draws of s from this distribution correspond to different interpretations of the same prompt [13]. The editor implicitly samples or deterministically selects a value of s as part of its internal computation, even if this variable is not explicitly encoded.

The mapping from latent semantics to image space can be described via a decoder

$$G_\theta(x, m, s, z) \in \mathbb{R}^{H \times W \times C} \tag{33}$$

such that

$$F_\theta(x, m, e(p), z) \approx G_\theta(x, m, s, z), \quad s \sim \pi(\cdot \mid p). \tag{34}$$

In this view, ambiguity in the prompt translates into randomness in s, which in turn induces variability in the edited image. The set of acceptable outputs $\mathcal{A}(x, m, p)$ can be identified with the support of the distribution of $G_\theta(x, m, s, z)$ when s and z vary according to their conditional distributions.

From a probabilistic standpoint, one can define an ambiguity-induced variance of the editing operator. Consider the conditional expectation

$$\mu(x, m, p) = \mathbb{E}\big[G_\theta(x, m, s, z) \mid p\big] \tag{35}$$

and the corresponding covariance operator

$$\Sigma(x, m, p) = \mathbb{E}\Big[\big(G_\theta(x, m, s, z) - \mu(x, m, p)\big) \otimes \big(G_\theta(x, m, s, z) - \mu(x, m, p)\big)\Big] \tag{36}$$

The diagonal entries of $\Sigma(x, m, p)$ measure the variability of each pixel under different interpretations of the prompt and noise realizations. Large variance inside the masked region may be acceptable for highly ambiguous prompts, whereas large variance outside the mask suggests instability in localization.

Ambiguity at the lexical or syntactic level can also be described combinatorially [14]. The prompt p is a sequence of tokens drawn from a vocabulary. The space of paraphrases within a bounded edit distance in token space forms a discrete neighborhood

$$\mathcal{N}_{\text{lex}}(p) = \big\{p' : d_{\text{edit}}(p, p') \le \kappa\big\} \tag{37}$$

for some integer threshold κ. Within this neighborhood, many prompts may be semantically equivalent or similar. Ambiguity can be formalized by considering equivalence classes of prompts that map to overlapping or identical distributions over latent semantics. The cardinality and structure of such equivalence classes are objects of discrete combinatorial interest and influence the effective dimensionality of the prompt space encountered during training.

Adversarial prompts are constructed with an objective function in mind. Given a base configuration (x, m, p), the adversary selects a modified prompt p' from a constraint set $\mathcal{S}(p)$ in order to maximize a loss on the editor's output. Let

$$\mathcal{L}(x, m, p', z; \theta) = L\big(F_\theta(x, m, e(p'), z); x, m, p\big) \tag{38}$$

be a task-dependent loss measuring mislocalization, semantic drift, or violation of constraints. The adversarial prompt search

Optimization space	Constraint set	Objective (schematic)	Notes
Discrete token space	$\mathcal{S}(p) = \{p' : d_{\text{edit}}(p, p') \leq \kappa\}$	$\max_{p' \in \mathcal{S}(p)} \mathbb{E}_z[\mathcal{L}(x, m, p', z; \theta)]$	Captures paraphrases and small token-level perturbations
Continuous embedding space	$\{\delta : \|\delta\| \leq \varepsilon\}$ around $e(p)$	$\max_{\|\delta\| \leq \varepsilon} \mathbb{E}_z[\mathcal{L}(x, m, e(p) + \delta, z; \theta)]$	Enables gradient-based prompt attacks in representation space
Semantics-constrained prompts	$\mathcal{S}(p)$ preserving coarse meaning while altering wording	Maximize mislocalization or semantic drift while remaining on-topic	Models subtle adversarial rephrasings used by human attackers
Bad-embedding region $\Omega_{\text{bad}}(x, m)$	$\{e : \mathbb{E}_z[D_{\text{out}}(x, F_\theta(x, m, e, z); m)] > \tau\}$	Characterize problematic embeddings that cause large leakage	Guides robust design, prompt filtering, and safe operating regimes

Table 8 Adversarial prompt formulations in token, embedding, and semantic spaces for analyzing worst-case robustness.

problem is [15]

$$p^{\text{adv}} \in \arg\max_{p' \in \mathcal{S}(p)} \mathbb{E}_z\big[\mathcal{L}(x, m, p', z; \theta)\big]. \tag{39}$$

In practice, the adversary may operate in the continuous embedding space. Treating the encoder $e(\cdot)$ as a differentiable mapping, one can consider perturbations of the embedding

$$e' = e(p) + \delta, \quad \|\delta\| \leq \varepsilon \tag{40}$$

for some radius ε. The adversarial perturbation problem in embedding space then becomes

$$\delta^\star \in \arg\max_{\|\delta\| \leq \varepsilon} \mathbb{E}_z\big[\mathcal{L}(x, m, e(p) + \delta, z; \theta)\big]. \tag{41}$$

Under differentiability assumptions, this problem can be attacked using gradient-based methods that approximate the direction of steepest ascent of the expected loss with respect to the embedding.

The relationship between discrete prompt manipulations and continuous embedding perturbations is governed by the geometry of the encoder. For small perturbations, a linear approximation yields

$$\mathcal{L}(x, m, e(p)+\delta, z; \theta) \approx \mathcal{L}(x, m, e(p), z; \theta) + \langle \mathbb{0}_e \mathcal{L}(x, m, e(p), z; \theta), \delta \rangle \tag{42}$$

Here, $\mathbb{0}_e \mathcal{L}$ is a gradient in embedding space, and its direction indicates how small changes in embedding affect the loss. If the gradient's projection onto directions that correspond to plausible textual perturbations is large, then the editor is sensitive to adversarial prompt modifications.

Ambiguity and adversariality are not mutually exclusive. An adversarial prompt may deliberately exploit ambiguous phrases that, when interpreted in a particular way by the model, produce undesired content [16]. In such cases, the ambiguity set $\mathcal{A}(x, m, p)$ contains both benign and harmful outputs. The adversary's goal is to bias the model's internal selection mechanism toward harmful regions of $\mathcal{A}(x, m, p)$. This can be conceptualized by introducing a parameterized family of distributions over latent semantics

$$\pi_\eta(s \mid p) \tag{43}$$

where η represents internal attention or conditioning parameters influenced by the adversarial prompt. Changes in η triggered by subtle differences in word choice can shift probability mass within $\mathcal{A}(x, m, p)$ from one subset of interpretations to another.

From a geometric perspective in embedding space, one can examine the level sets of the loss function associated with robustness violations. Define a region

$$\Omega_{\text{bad}}(x, m) = \big\{e : \mathbb{E}_z\big[D_{\text{out}}(x, F_\theta(x, m, e, z); m)\big] > \tau\big\} \tag{44}$$

for a threshold τ. This region contains embeddings that tend to cause excessive modifications outside the mask. The robustness of the editor with respect to prompt embeddings can be analyzed by examining how large the intersection of $\Omega_{\text{bad}}(x, m)$ is with the manifold of embeddings achievable by realistic prompts. If this intersection occupies a significant fraction of the manifold's volume, then randomly sampled or slightly perturbed prompts may often lead to mislocalization.

In summary, modeling ambiguous and adversarial prompts requires tools from probability, discrete mathematics, and differential geometry of embedding spaces. Ambiguity is captured by distributions over latent semantics yielding sets of acceptable outputs, while adversariality is modeled as optimization over prompts or embeddings to maximize robustness-related losses. The interplay between these aspects determines how a mask-conditioned editor behaves under challenging prompt conditions [17].

5. Empirical Analysis of Robustness and Failure Modes

Empirical analysis of robustness for mask-conditioned image editors involves systematically probing the behavior of the model across a variety of prompts, masks, and images. While detailed numerical results depend on the particular architecture and dataset, the conceptual structure of such an analysis can be discussed using the metrics introduced earlier. The objective is to identify characteristic failure modes, quantify their frequency and severity, and relate them to properties of prompts and masks.

A natural starting point is to examine the distribution of the unmasked distance D_{out} across a representative dataset. For each sample (x, m, p) and noise draw z, one computes

$$\mathbb{F} = F_\theta(x, m, e(p), z) \tag{45}$$

and evaluates $D_{\text{out}}(x, \mathbb{F}; m)$. Aggregating these values yields an empirical distribution whose tail behavior indicates the prevalence of large unintended changes outside the mask. Robustness with respect to benign prompts can be summarized by statistics such as empirical mean, variance, and selected quantiles of D_{out}. Multivariate statistics are also informative: for instance, one can examine correlations between D_{out} and mask properties such as area, perimeter, or shape complexity.

Mask geometry can be characterized using discrete geometric descriptors. Let $M(m)$ be the set of masked pixels and consider the induced boundary

$$\partial M = \big\{(i, j) \in M(m) : \hbar(i', j') \in V \backslash M(m) \text{ adjacent to } (i, j)\big\}. \tag{46}$$

The length of this boundary, in terms of pixel adjacency, is $|\partial M|$ and reflects how much interface exists between edited and non-edited regions. Empirically, larger boundary lengths often correlate with increased risk of leakage, since convolutional and attention mechanisms propagate information across neighboring regions. By regressing observed D_{out} values on $|M(m)|$, $|\partial M|$, and other descriptors, one can gain insight into how localization robustness depends on mask complexity.

Ambiguous prompts can be studied by constructing sets of prompts that are intended to describe similar content with varying degrees of specificity [18]. For each base description, multiple paraphrases can be generated, and the editor's outputs compared. Suppose a set of prompts $\{p, \ldots, p_K\}$ is designed to share approximately the same semantics for a fixed (x, m). For each p_k, one can sample outputs $\mathbb{F}_{k,\ell}$ for several noise draws z_ℓ. The variability across k at fixed ℓ, and across ℓ at fixed k, reflects different components of randomness. Statistical decompositions such as analysis of variance can separate the contribution of prompt variation from stochastic noise in the generative process.

Semantic robustness can be evaluated using embedding-based alignment scores. For each edited image $\mathbb{F}$ and its prompt p, semantic similarity $S_{\text{in}}(\mathbb{F}, p; m)$ can be computed. Under ambiguous prompts, the goal is not necessarily to maximize this similarity to a single canonical embedding, but to remain within an acceptable range for at least one interpretation. If a set of reference embeddings $U(x, m, p)$ is available, perhaps obtained from multiple human annotators who provide alternative verbal interpretations, then the ambiguity-aware loss $L_{\text{amb}}(\mathbb{F}; x, m, p)$ can be estimated. Observing how this quantity varies with prompt structure, such as the presence of metaphors or rare adjectives, reveals how the editor handles linguistic complexity.

Adversarial prompt analysis requires constructing or approximating solutions to the optimization problem defining p^{adv} or $\delta^\star$. In a white-box setting where gradients with respect to the prompt embedding are accessible, one can iteratively update the embedding using gradient ascent while projecting back onto the manifold of embeddings that correspond to valid token sequences. In a black-box setting, finite-difference approximations or evolutionary strategies can be used. Once adversarial prompts are obtained, their impact is evaluated using robustness metrics such as D_{out}, $C(x, \mathbb{F}; m)$, and semantic drift scores.

Empirically, several recurring failure modes are observed in mask-conditioned editing [19]. One frequent pattern is halo artifacts or color bleeding at the mask boundary, where the editor adjusts pixels immediately outside the mask to maintain local coherence. This effect becomes more pronounced for high-contrast edits or when the requested content significantly alters lighting or shadows. Another pattern is global style transfer, where a stylistic prompt causes the model to modify the overall color palette or texture of the image, even in regions outside the mask. This often arises when the prompt includes global style descriptors such as artistic movement names or global color schemes.

A distinct failure mode appears under complex or ambiguous prompts that combine multiple attributes and objects. In such cases, the editor may misattribute certain adjectives or modifiers to background objects rather than the intended masked region. This kind of semantic mislocalization is difficult to quantify purely with pixel-wise metrics, but embedding-based analyses can detect it when semantic content associated with the prompt appears in unmasked regions in the image embedding.

Robustness under adversarial prompts exhibits its own patterns. Carefully crafted prompts can cause the mask to be effectively ignored, leading to global replacements or scene alterations akin to unconditional text-to-image generation starting from the input as a loose reference. Others can exploit weaknesses in safety filters: for instance, by embedding disallowed content descriptions within benign-sounding phrases, or by using rare or coded terms that the filter has not been trained to recognize [20]. In embedding space, such adversarial prompts often occupy directions of large gradient magnitude for losses associated with safety or localization.

An interesting question from numerical analysis is how the discretization of the generative process affects robustness. In diffusion-based editors, the generative process is implemented as a discretized stochastic differential equation or ordinary differential equation over a sequence of time steps. Let the process be represented as a sequence of states

$$h_\varkappa, h, \ldots, h_T \tag{47}$$

with update rule

$$h_{t+} = \Psi_\theta(h_t, x, m, e(p), \eta_t) \tag{48}$$

where Ψ_θ is a learned update function and η_t are random noise terms. Step size selection, number of steps T, and numerical stability of Ψ_θ all influence how errors accumulate. Smaller step sizes may reduce the amplification of prompt perturbations, but also increase computational cost. Viewing the process as a numerical integrator, robustness can be connected to stability regions of the implicit dynamical system and to Lipschitz properties of Ψ_θ with respect to $e(p)$.

Empirical analysis thus provides a bridge between the abstract robustness metrics and observable failure modes. By systematically measuring distances, similarities, and variances

across diverse prompts and masks, one can quantify the extent of undesired behavior and identify conditions under which mask-conditioned editors are more or less reliable.

6. Mitigation Strategies and Theoretical Insights

Mitigating robustness issues for mask-conditioned image editors involves modifying training objectives, architectural components, or inference procedures to reduce sensitivity to ambiguous and adversarial prompts while preserving editing expressiveness. Theoretical insights from linear algebra, optimization, and probability can guide the design of such strategies [21].

One approach is to explicitly penalize changes outside the mask during training. Suppose training data includes pairs $(x, m, p, x^\star)$ where $x^\star$ is a target edited image. A standard reconstruction loss might focus on the masked region. To encourage background preservation, one can add a term

$$L_{\text{bg}}(\theta) = \mathbb{E}\Big[\big\|(x - F_\theta(x, m, e(p), z)) \odot (\,-m)\big\|\Big] \tag{49}$$

to the training objective. This term is quadratic in the difference between background pixels, resembling an ℓ penalty. From a linear algebra perspective, this acts as a projection of the error onto the subspace spanned by unmasked pixels, and penalizing its squared norm constrains the component of the model's output that deviates from the identity transformation on that subspace.

Another strategy involves regularizing the Jacobian of the editor with respect to the prompt embedding, especially on the unmasked region. As discussed earlier, the Jacobian

$$J_{\text{out}}(x, m, e, z) = P_{\text{out}} \circ \mathbb{0}_e F_\theta(x, m, e, z) \tag{50}$$

captures local sensitivity of the background to changes in the embedding. A Jacobian regularization term can be introduced,

$$L_{\text{jac}}(\theta) = \mathbb{E}\big[\|J_{\text{out}}(x, m, e(p), z)\|_F\big] \tag{51}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Minimizing L_{jac} encourages the mapping from embedding to background pixels to be locally flat, thus reducing vulnerability to small adversarial perturbations in embedding space. Theoretically, if the operator norm of J_{out} is bounded by a constant L, then for all perturbations δ within a ball of radius ε, one has

$$\big\|P_{\text{out}}\big[F_\theta(x, m, e(p) + \delta, z) - F_\theta(x, m, e(p), z)\big]\big\| \le L\varepsilon. \tag{52}$$

This Lipschitz-type bound provides a quantitative guarantee on local robustness [22].

Ambiguity-aware training can be used to handle prompts with multiple acceptable interpretations. Instead of associating a single target $x^\star$ with each ambiguous prompt, one can work with a set of targets or a distribution over them. The loss function becomes

$$L_{\text{amb}}(\theta) = \mathbb{E}\Big[\inf_{x^\star \in \mathcal{A}(x,m,p)} \ell\big(F_\theta(x, m, e(p), z), x^\star\big)\Big] \tag{53}$$

where ℓ is a reconstruction or perceptual loss. This formulation recognizes that multiple outputs may be acceptable and trains the editor to stay near at least one acceptable mode for each ambiguous prompt. From an optimization viewpoint, this introduces a nonconvex infimum inside the expectation, which may be approximated by sampling or by using surrogate losses that capture distance to a convex hull of acceptable targets.

Robust optimization techniques can also be applied in prompt space. Given a set of perturbations $\mathcal{S}(p)$ around each prompt, the training objective can be modified to minimize adversarial risk,

$$L_{\text{rob}}(\theta) = \mathbb{E}\Big[\sup_{p' \in \mathcal{S}(p)} L\big(F_\theta(x, m, e(p'), z), x^\star\big)\Big]. \tag{54}$$

In practice, the supremum is approximated by generating adversarial prompts during training via gradient-based updates in embedding space or via search over token-level perturbations. This adversarial training procedure encourages the editor to learn parameters that are stable under a neighborhood of prompts around those seen in data, thereby improving robustness to both benign and malicious variations.

Architecturally, mask injection mechanisms can be adjusted to better enforce locality [23]. Many editors inject the mask into intermediate feature maps by concatenation or masking operations. One can strengthen spatial decoupling by introducing explicit block-diagonal structures in linear layers or attention patterns that reduce interactions between masked and unmasked regions. For example, consider a linear layer with weight matrix W mapping flattened pixel features to a latent representation. If the mask induces a partition of pixels into two groups, one can approximate W by a block matrix

$$W \approx \begin{pmatrix} W_{\text{in}} & \varkappa \\ \varkappa & W_{\text{out}} \end{pmatrix} \tag{55}$$

so that masked and unmasked regions do not directly interact at that layer. While exact block-diagonality may be too restrictive, penalizing the off-diagonal blocks via a term

$$L_{\text{block}}(\theta) = \|W_{\text{cross}}\|_F \tag{56}$$

where W_{cross} collects cross-region weights, can reduce leakage.

Attention mechanisms conditioned on the prompt embedding can also be regularized. In cross-attention layers, keys and values are derived from image features while queries come from text embeddings, or vice versa. One can enforce that attention weights associated with background pixels integrate to smaller values when the prompt is primarily about the masked region. This suggests adding constraints of the form [24]

$$\sum_{(i,j) \in V \setminus M(m)} \alpha_{ij} \le \gamma \tag{57}$$

for attention weights α_{ij} and a small threshold γ. In differentiable form, penalties on these sums can be incorporated in the training loss, shaping the attention distribution.

From a probabilistic standpoint, uncertainty estimates can be leveraged to inform users about robustness. If the variability of outputs under prompt rephrasings or noise realizations is

high, the editor could indicate that the requested operation is inherently ambiguous. One way to quantify such variability is through Monte Carlo sampling: for fixed (x, m, p), generate multiple outputs $\mathbb{F}, \ldots, \mathbb{F}_N$ and compute empirical variance in masked and unmasked regions. High variance in the unmasked region warns of potential robustness issues. While this does not directly mitigate the problem, it gives users feedback and can guide interface designs where the editor suggests more precise prompts or mask adjustments.

The interplay between robustness and expressiveness can be framed using bias–variance trade-offs. Aggressive regularization of the Jacobian or strong penalties on background changes may restrict the model's ability to perform radical edits that require coordinated global adjustments, such as changing lighting or inserting large objects. Linear algebraic analyses of the model's representation capacity under such constraints can provide insight into what classes of edits remain achievable. For example, by modeling the editor as an approximately linear operator in a neighborhood of an image, one can study the subspace of directions in image space that remain accessible under strong locality constraints.

Finally, discrete prompt manipulations and token-level defenses play a complementary role [25]. By analyzing which tokens or token combinations frequently appear in adversarial prompts that cause misbehavior, one can construct filters or remapping strategies that neutralize such patterns. From a discrete mathematics viewpoint, this corresponds to identifying subgraphs in the token transition graph that correlate with robustness failures and adjusting language processing pipelines to downweight or reinterpret such transitions.

These mitigation strategies, grounded in mathematical considerations, illustrate how robustness of mask-conditioned image editors to ambiguous and adversarial prompts can be improved. They highlight trade-offs between local invariance, global coherence, and semantic flexibility, and suggest directions for future work on principled training and architectural modifications.

7. Conclusion

Mask-conditioned image editors provide a flexible interface for localized, prompt-driven image manipulation, but their behavior under ambiguous and adversarial prompts raises important questions about robustness, predictability, and control. This paper formulated a general mathematical model of such editors as stochastic operators acting on image tensors, binary masks, and prompt embeddings. Within this framework, several robustness metrics were introduced to capture preservation of the unmasked region, localization of changes inside the mask, and semantic alignment between edited content and textual instructions.

Ambiguous prompts were modeled through distributions over latent semantic variables and sets of acceptable outputs, highlighting how under-specified language can induce variability in the generated images. Adversarial prompts were conceptualized as solutions to optimization problems in discrete token space or continuous embedding space, aiming to maximize losses associated with mislocalization or semantic drift. The analysis connected these ideas to Jacobian-based sensitivity measures and to geometric properties of the prompt embedding manifold [26].

Empirical considerations were discussed in terms of how robustness metrics are observed to behave across datasets of prompts, masks, and images. Characteristic failure modes such as leakage of edits beyond the mask, global style shifts, and semantic mislocalization were related to mask geometry, attention patterns, and numerical aspects of generative inference. Although specific quantitative results depend on particular architectures and datasets, the structural features of these failure modes appear to reflect general properties of mask-conditioned editing rather than isolated implementation details.

Mitigation strategies were examined from both practical and theoretical perspectives. Mask-aware loss terms, Jacobian regularization, ambiguity-aware training objectives, and robust optimization formulations in prompt space were proposed as ways to constrain the editor's sensitivity to prompt variations. Architectural adjustments, including block-structured transformations and regularized attention, were discussed as mechanisms for limiting cross-region interactions. These methods involve trade-offs that shape the space of edits the model can perform while maintaining localization and stability. The treatment emphasized a neutral analysis of robustness rather than prescriptive conclusions. The mathematical formulations and conceptual tools presented here are intended to support further empirical and theoretical work on the reliability of mask-conditioned image editors under complex prompting behavior. Future research may refine these models, develop more precise bounds on robustness, and explore user interface designs that help mitigate ambiguity and guide safe, predictable editing interactions [27].

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] Z. Wang, B. Cai, Y. L. Wang, G. Y. Lv, R. C. Shan, and M. Zhang, "Rcar - a comparison study on the adaptive scale estimation of correlation filter-based visual tracking methods," *Robotics and biomimetics*, vol. 4, pp. 11–11, 11 2017.

[2] J. Y. Lee, J. Jeong, E. M. Song, C. Ha, H. J. Lee, J. E. Koo, D.-H. Yang, N. Kim, and J.-S. Byeon, "Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets.," *Scientific reports*, vol. 10, pp. 8379–8379, 5 2020.

[3] A. Revanur, D. D. Basu, S. Agrawal, D. Agarwal, and D. Pai, "Mask conditioned image transformation based on a text prompt," Nov. 21 2024. US Patent App. 18/319,808.

[4] O. Kaynak, "The golden age of artificial intelligence," *Discover Artificial Intelligence*, vol. 1, pp. 1–7, 9 2021.

[5] Y. Yang, "The potential energy of artificial intelligence technology in university education reform from the per-

spective of communication science," *Mobile Information Systems*, vol. 2021, pp. 1–7, 9 2021.

[6] R. V. Bidwe, S. Mishra, S. Patil, K. Shaw, D. R. Vora, K. Kotecha, and B. Zope, "Deep learning approaches for video compression: A bibliometric analysis," *Big Data and Cognitive Computing*, vol. 6, pp. 44–44, 4 2022.

[7] Y. Han, Z. Zhang, F. Chen, and K. Chen, "An efficient approach for shadow detection based on gaussian mixture model," *Journal of Central South University*, vol. 21, pp. 1385–1395, 4 2014.

[8] M. U. R. Siddiqi, W. Ijomah, G. Dobie, M. Hafeez, S. G. Pierce, W. Ion, C. Mineo, and C. MacLeod, "Low cost three-dimensional virtual model construction for remanufacturing industry," *Journal of Remanufacturing*, vol. 9, pp. 129–139, 10 2018.

[9] A. Sethi, M. Rahurkar, and T. S. Huang, "Event detection using variable module graphs for home care applications," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 111–111, 4 2007.

[10] S. Zhang, "Paraviewweb architecture method of power security emergency drill platform based on vr technology," *Multimedia Tools and Applications*, vol. 83, pp. 6447–6467, 6 2023.

[11] H. Zhou, S. Li, A. Li, Q. Huang, F. Xiong, N. Li, J. Han, H. Kang, Y. Chen, Y. Li, H. Lin, Y.-H. Zhang, X. Lv, X. Liu, H. Gong, Q. Luo, S. Zeng, and T. Quan, "Gtree: an open-source tool for dense reconstruction of brain-wide neuronal population.," *Neuroinformatics*, vol. 19, pp. 305–317, 8 2020.

[12] B. Natarajan and R. Elakkiya, "Dynamic gan for high-quality sign language video generation from skeletal poses using generative adversarial networks," 8 2021.

[13] D. Aleksandar, P. Ivan, and N. Mirjana, *Digital Performing Arts - Participatory Practices in a Digital Age*. 3 2023.

[14] Z. Szűts, "A critical approach to digital pedagogy – the search for an organic methodology in the information society," *Opus et Educatio*, vol. 6, 12 2019.

[15] T. Kaluarachchi, A. Reis, and S. Nanayakkara, "A review of recent deep learning approaches in human-centered machine learning.," *Sensors (Basel, Switzerland)*, vol. 21, pp. 2514–, 4 2021.

[16] S. Wirth, *Interfaces of AI*, pp. 217–234. transcript Verlag, 12 2023.

[17] R. Kestur, A. Angural, B. Bashir, S. N. Omkar, G. Anand, and M. B. Meenavathi, "Tree crown detection, delineation and counting in uav remote sensed images: A neural network based spectral–spatial method," *Journal of the Indian Society of Remote Sensing*, vol. 46, pp. 991–1004, 6 2018.

[18] B. J. Antony, S. Maetschke, and R. Garnavi, "Automated summarisation of sdoct volumes using deep learning: Transfer learning vs de novo trained networks.," *PloS one*, vol. 14, pp. e0203726–, 5 2019.

[19] G. Zhu, Z. Qu, L. Sun, Y. Liu, and J. Yang, "Realistic real-time processing of anime portraits based on generative adversarial networks," *Journal of Real-Time Image Processing*, vol. 21, 6 2024.

[20] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan, "Acm multimedia - near-duplicate keyframe retrieval by nonrigid image matching," in *Proceedings of the 16th ACM international conference on Multimedia*, pp. 41–50, ACM, 10 2008.

[21] S. S. A. Ghadani* and D. C. Jayakumari, "Innovating fire detection system fire using artificial intelligence by image processing," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, pp. 349–356, 9 2020.

[22] A. Budhiraja, "Image cartoonification using machine learning: Transforming visual content with artistic intelligence," *The Pharma Innovation*, vol. 8, pp. 701–706, 1 2019.

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

[24] K. Gajjar, T. van Niekerk, T. D.-I. Wilm, and P. Mercorelli, "Vision-based deep learning algorithm for detecting potholes," 2 2021.

[25] Y. Dong and W. D. Pan, "A survey on compression domain image and video data processing and analysis techniques," *Information*, vol. 14, pp. 184–184, 3 2023.

[26] G. A. Mitchell, "Distinctive expertise: Multimedia, the library, and the term paper of the future," *Information Technology and Libraries*, vol. 24, pp. 32–36, 3 2005.

[27] R. Rosas-Romero, O. Lopez-Rincon, and O. Starostenko, "Fully automatic alpha matte extraction using artificial neural networks," *Neural Computing and Applications*, vol. 32, pp. 6843–6855, 3 2019.