Analyzing Corrupted Segmentation Maps in ControlNet: A Study of Robustness and Reliability-Aware Conditioning

Sanjana Duttagupta
Massachusetts Institute of Technology
sanjd@mit.edu

Abstract

Diffusion-based image generation models such as ControlNet enable fine-grained control over generated outputs through structured conditions inputs such as segmentation maps. However, these methods implicitly assume that inputs are clean and accurate, which limits their robustness in real-world settings where they may be noisy or imperfect. This work investigates the impact of corrupted segmentation map inputs on image generation using a pretrained ControlNet with a frozen Stable Diffusion backbone and evaluates a simple reliability-aware conditioning strategy to mitigate performance degradation.

Realistic annotation noise is simulated by applying boundary perturbations, region dropout, and label flipping to segmentation maps obtained from the ADE20K dataset. To estimate input quality, a lightweight reliability metric based on edge density and structural fragmentation is introduced to adaptively scale the influence of the pretrained segmentation conditioning branch during inference.

The experiments show that segmentation noise significantly degrades generation quality, as measured by CLIP similarity, particularly in the case of label flipping which caused a semantic degradation of up to 32% in the baseline. The reliability-aware approach consistently improves performance at moderate noise levels, achieving a peak improvement of 18.0% in CLIP similarity under 50% label flip noise. While performance gains diminish at extreme noise levels, the system consistently shows better performance at high noise levels with the reliability estimator versus without. Overall, this work demonstrates that even simple, non-learned reliability estimation can improve the robustness of conditional diffusion models, demonstrating the importance of incorporating input quality awareness into generative pipelines.

The code for this project is available at: https://colab.research.google.com/drive/1nLf1uNOCy9Dj-f-7cRmzeia4tpKE_OjY?usp=sharing

1. Introduction

1.1. Research Problem

Generative Artificial Intelligence (GenAI) has made strides in generating media ranging from hyper-realistic images to detailed text and even molecular design [3]. In particular, generative imaging has transformed computer vision through diffusion models, which construct images by refining random noise via a reverse diffusion process [3]. Building on this, Stable Diffusion is a text-to-image model that improves efficiency by operating in a latent space (machine-learned representations of large media datasets) [7, 9]. For greater control over image generation, Zhang et al. [10] introduced ControlNet, a neural network architecture that augments Stable Diffusion with an additional trainable branch. This branch enables the incorporation of conditional inputs which are structure-defining inputs like edge maps, depth maps, and segmentation maps to guide the structure of the generated image [10]. The model takes both a text prompt and a conditioning signal which allows the users to influence not only the content of the generated image but also the structure. While effective, ControlNet relies on well-curated conditioning inputs, and like many learning systems, is sensitive to noise in the input [5]. An example of a conditioning input is segmentation maps, which are used to represent object structure and spatial layout using distinct colors and boundaries [4]. Due to variability in human labeling and image quality, segmentation maps often contain noise, including imprecise boundaries, missing regions, and artifacts [4]. Although prior work has shown segmentation maps' effectiveness as conditioning signals, it typically assumes these inputs are reliable and clean [10].

1.2. Motivation

Noisy or ambiguous segmentation maps as inputs to the ControlNet architecture have been largely underexplored. This demonstrates a limitation in current conditional generative pipelines since noisy inputs can lead to outputs that deviate from user intent. This becomes especially important as diffusion-based models are increasingly integrated into

real-world systems such as design tools and medical imaging workflows, where robustness and reliability are essential [1, 7]. By analyzing how ControlNet performance degrades under noisy conditions, we can identify under which types of corruption the model begins to fail or produce semantically inconsistent outputs. This allows us to characterize the robustness boundaries of the model rather than assuming uniform performance across input quality levels.

By studying these failure modes, we can move beyond qualitative observations and develop a more principled understanding of sensitivity to input perturbations. In particular, introducing a reliability estimation mechanism alongside the generative pipeline enables the model to assess the trustworthiness of a given segmentation map before conditioning on it. Such an estimator could predict when an input is likely to produce unstable or misleading generations, allowing downstream systems to down-weight corrupted conditions. This is especially valuable in domains where even small deviations from intended structure can have significant consequences.

The goal of this research is to evaluate the structural and semantic robustness of the ControlNet architecture under corrupted segmentation inputs. This study will consider various forms of possible noise in segmentation maps across different levels of noise and analyze the resulting generated image. Beyond identifying these failure modes, this work tests and validates a simple reliability-aware conditioning framework. The framework is designed to improve the model's robustness by scaling the influence of the ControlNet branch based on a calculated reliability score Q. The research intends to demonstrate that incorporating a self-skepticism mechanism into conditional diffusion pipelines can mitigate the negative impacts of imperfect real-world data without requiring additional model training or fine-tuning. This ensures a low barrier of entry to implementing a reliability-estimator and maintains within the computation limitations of this project.

1.3. Related Works

Previous work has similarly demonstrated efforts to extend the flexibility and capabilities of ControlNet. For instance, LooseControl improves the quality of generated outputs by relaxing strict conditioning constraints for generalized depth maps, allowing for more flexible alignment between the conditioning signals and structure generated in the images [2]. Similarly, DC-ControlNet extends the framework to multi-condition image generation, enabling outputs to better reflect complex user intent through inter- and intra-element conditioning [8].

These methods differ from this research because they do not explicitly address that conditioning signals are often noisy or incomplete. Existing approaches assume that conditioning inputs are reliable, or that flexibility in the conditioning mechanism can handle variation in input quality, but robustness to corruption is not directly modeled. This assumption is limiting because increasing flexibility or adding additional conditioning channels does not inherently improve robustness to noisy inputs. In fact, it may amplify failure modes by causing the model to rely more heavily on flawed conditioning signals, especially when those signals are still partially structured or visually plausible.

In contrast, this research focuses directly on the robustness of ControlNet under imperfect conditioning. Rather than introducing additional structural complexity, this research analyzes how noise in segmentation maps affects generation quality and identifies the regimes under which ControlNet begins to fail. A lightweight reliability-aware is introduced to allow the model to adjust its reliance on conditioning signals based on their estimated quality.

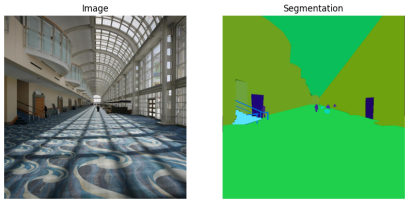
This is especially important for real-world deployment settings, where perfect annotations cannot be assumed. By enabling the model to distinguish between reliable and unreliable conditioning, this approach improves both generation stability and alignment with user intent under noisy conditions without introducing significant computational overhead.

1.4. Technical Approach Overview

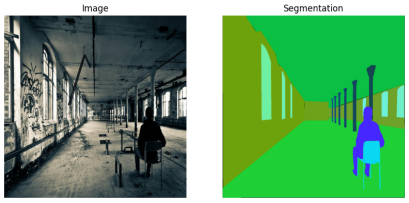
This research investigates how segmentation map corruption affects conditional image generation in ControlNet, and whether robustness can be improved without modifying the underlying diffusion model. A frozen pretrained ControlNet is built on a Stable Diffusion backbone, ensuring that all observed effects are attributable solely to changes in the conditioning input rather than model adaptation. Experiments are conducted on a filtered subset of the ADE20K dataset focused on a single semantic class to maintain consistency across generations. Samples of the input images and their corresponding segmentation maps obtained from ADE20K are shown in Figure 1.

To study robustness, realistic segmentation noise is implemented through three corruption types which are each applied at multiple levels. For each corrupted segmentation map, we generate images using a fixed text prompt and controlled sampling setup, enabling direct comparison against clean ground-truth conditions. We then evaluate how output quality degrades under increasing corruption using both semantic alignment and structural consistency metrics.

Finally, a simple reliability-aware mechanism is used to estimate a quality score from observable segmentation statistics and uses it to tune the strength of ControlNet conditioning at inference time. This allows the model to dynamically reduce reliance on low-quality inputs and shift emphasis toward the text prior when segmentation signals are unreliable, without requiring any retraining or architectural changes.



(a) Sample 1



(b) Sample 2

(c) Sample 3

Figure 1. **Samples from the ADE20K dataset used in this study.** Images that were collected contain a "chair" label to ensure semantic consistency across conditioning trials. Each sample contains a natural image and its corresponding segmentation map, which is the structural conditioning input for ControlNet. These examples show the type of scene layouts used in the experiments, where chairs appear in different configurations. This variability is important for evaluating robustness, as it ensures the model is tested on both simple and cluttered indoor structures.

2. Methods

2.1. Research Objective

This research considers how noise in segmentation maps as a conditional input to ControlNet influence generated images. Although these models rely on clean inputs, real-world segmentation map images are often imperfect due to subjectivity in human labeling and variation in image quality [4]. The goal is to explore whether the performance of ControlNet under imperfect conditions can be improved without modifying the underlying model. The research focuses on three main objectives: (i) measuring how different types and levels of segmentation corruption affect generated images, (ii) analyzing common mode of failures that appear under noisy conditions, and (iii) evaluating a simple reliability-based adjustment framework designed to improve robustness.

2.2. Method Overview

This study uses a pretrained ControlNet model with a frozen Stable Diffusion backbone [10] with no additional training or fine-tuning to to ensure the isolation of the effect of the conditioning inputs only. The experiments are conducted with a subset of the ADE20K dataset [11, 12], specifically filtered to images containing the "chair" class to maintain semantic consistency during testing. The study compares the outputs generated from clean/ground-truth inputs and perturbed inputs, where the perturbed inputs are corrupted versions of the ground-truth segmentation with either boundary noise, region dropout, or label flipping at various noise levels. A lightweight reliability estimator is introduced to address the observed sensitivity to noisy conditioning. The estimator computes a reliability score Q from observable segmentation characteristics, including edge density and fragmentation,

2.3. Segmentation Corruption Process

To simulate realistic imperfections, corruption functions are applied to clean segmentation maps S_{clean}, producing noisy versions $\tilde{S}$. The corruptions reflect three common failure cases:

- **Boundary perturbation**: Introduces spatial noise along object edges to simulate imperfect manual labeling.
- **Region dropout**: Replaces individual pixels with random semantic values across the image, with a probability determined by the noise level. This simulates high-frequency noise in the segmentation map.
- **Label flipping**: Replaces semantic labels with incorrect classes with a probability determined by the noise level. This allows us to simulate semantic noise in our inputted image.

Each corruption type is evaluated across four intensity levels: 0%, 25%, 50%, and 75%. These levels serve two purposes: first, to observe the degradation of the model's output under noisy conditions, and second, to stress-test the reliability estimator. This allows us to verify if the proposed Q-metric can dynamically recognize when to attenuate the conditioning signal in favor of the prompt.

2.4. Generation Pipeline

For each segmentation map, we use a fixed text prompt and generate images through the ControlNet-conditioned diffusion process. To ensure that differences in output are attributable to segmentation quality only, all external parameters (including random seeds, text prompts, and sampling steps) are held constant across all conditions.

2.5. Reliability-Aware Conditioning

To mitigate the effect of noisy inputs, we introduce a reliability score $Q \in [0, 1]$. This score is computed from observable corruption statistics: edge density (proportion of

boundary pixels) and fragmentation (number of connected components). During inference, Q is used to adaptively scale the ControlNet conditioning strength. The intuition is that lower-quality segmentation maps should exert less influence on the final generation, allowing the model to rely more heavily on the text prompt priors. The full pipeline is shown in Figure 2.

2.6. Performance Metrics

To evaluate performance, we employ both semantic and structural criteria.

CLIP Similarity. Contrastive Language-Image Pre-training (CLIP) [6] is used to measure semantic alignment between the generated image and the text prompt. CLIP utilizes a dual-encoder architecture to project both images and text into a shared embedding space. Higher scores indicate that the generated image better adheres to the text prompt despite corrupted conditioning.

Structural Consistency. To evaluate structural fidelity, we define a Boundary Alignment Score (BAS). This metric uses edge detection operators (the Sobel filter for segmentation maps and the Canny operator for generated images) to extract binary edge masks (images where pixels are assigned a 0 or 1 depending on whether a pixel is part of an edge or the background). These masks are compared via an Intersection-over-Union (IoU) calculation:

$$\text{BAS} = \frac{|E_{seg} \cap E_{img}|}{|E_{seg} \cup E_{img}|}$$

where E_{seg} and E_{img} represent the spatial boundaries of the input and output, respectively. This quantitative measure captures the degree to which the model sticks to the structure of the conditioning inputs.

2.7. Experimental Design

The study utilizes two experimental conditions to isolate the benefit of the reliability estimator:

- **Baseline:** Generation using corrupted segmentation maps with noise levels ranging from 0.0 to 0.75.
- **Reliability-Adjusted:** Generation using corrupted maps where the conditioning strength is modulated by the predicted Q score.

Comparing the Baseline to the Reliability-Adjusted condition allows us to compare the quality of the image with adaptive weighting versus without.

3. Results

3.1. Experimental Setup

The robustness of ControlNet under segmentation noise is evaluated using a subset of ADE20K. To ensure results are not biased by specific image geometries, $n = 20$ samples are evaluated for each noise condition. The mean CLIP score and mean structural alignment are collected across these samples to provide a stable measure of model performance at each noise level. These samples are filtered and are only selected if the label "chair" is contained within their segmentation map to ensure semantic consistency across the selected samples. The text prompt allows for both structural alignment to the segmentation map and alignment to additional elements in the prompt by instructing the generated image to contain chairs in addition to other objects and styles:

> **Experimental Text Prompt (T):**
> *"cozy indoor room with chairs, table, and soft ambient lighting"*

We subject the model to three distinct failure modes: boundary noise, region dropout, and label flipping, across four intensity levels $\{0.0, 0.25, 0.5, 0.75\}$ as described in the Methods section. All images are generated with a fixed random seed and text prompt to isolate the effect of input corruption.

3.2. Results Overview

3.2.1. Quantitative Results

Analysis of Table 1 reveals that ControlNet's sensitivity is dependent on the noise type. Boundary Noise had a negligible impact on semantic scores, suggesting that the diffusion backbone inherently smooths edge-level imprecision and is therefore surprisingly robust to low-resolution masks. In contrast, Region Dropout and Label Flipping caused substantial semantic degradation, reducing CLIP similarity scores by up to 32% in the baseline configuration. These show that while the model can tolerate inaccurate boundaries, it struggles to recover from corrupted labels or missing regions, which demonstrate that the model is highly affected by noise that impacts the semantic understanding of objects in the segmentation map.

The effectiveness of the reliability estimator is demonstrated in Table 2. The largest improvement occurs under 50% Label Flip noise, where the reliability-aware adjustment improves CLIP similarity by 18.0%. This suggests that the model is no longer over-adhering to corrupted conditioning inputs. By incorporating the reliability estimator Q, the framework dynamically increases the influence of the text prompt while reducing dependence on unreliable segmentation information during image generation.

3.2.2. Qualitative Results

These quantitative trends are further supported qualitatively in Figure 3, which shows the generated outputs for Sample 3 from Figure 1. The reliability-aware method (bottom row) avoids many of the structural hallucinations visible in the baseline outputs. As noise severity increases, the

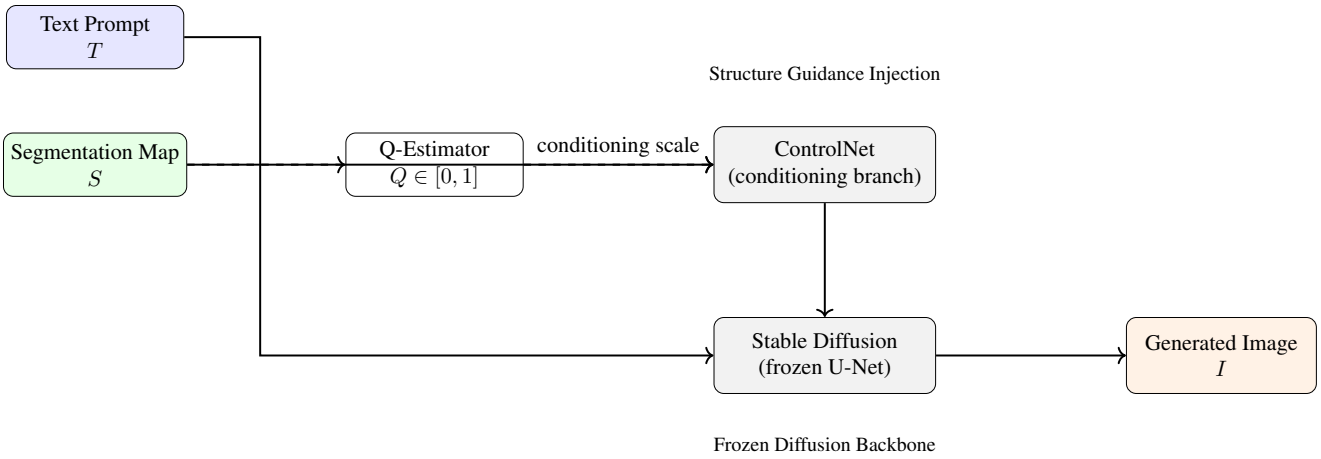


Figure 2. **Overview of the proposed reliability-aware ControlNet pipeline**. Standard ControlNet architectures assume that conditioning inputs, such as segmentation maps, are accurate and structurally consistent but these inputs are often noisy in real-world settings, leading to deviations from user intent. To address this limitation, a text prompt and segmentation map are provided to a frozen Stable Diffusion backbone through a ControlNet conditioning branch, while a lightweight reliability estimator predicts a quality score (Q) from observable properties of the segmentation map. This score is used to modulate the conditioning strength during inference, allowing the model to reduce reliance on unreliable segmentation inputs and place greater emphasis on the learned text prior. The proposed framework improves robustness to corrupted conditioning signals without requiring retraining of the underlying diffusion model.

baseline generations become increasingly incoherent, and the reliability-adjusted outputs remain visually cleaner and more semantically aligned with the prompt. At extreme noise levels (0.75), the segmentation entropy becomes too high for the Q-metric to fully recover realistic room structure. However, even in the Label Flip + Reliability case, the generated image still maintains prompt-consistent elements such as a chair and ambient lighting, demonstrating that the reliability-aware framework successfully shifts conditioning emphasis toward textual guidance when segmentation inputs become unreliable.

Further analysis of these results are explored in the next section.

3.3. Controlled Comparison Analysis

The ControlNet conditioning signal is considered across varying noise regimes to analyze the degradation in generated image quality with imperfect conditions.

3.3.1. Observation 1: Geometric vs. Semantic Noise

The model shows resilience to boundary noise. The implication here is that ControlNet operates as a high-pass filter in its early layers. It extracts the general gist of the structure while the diffusion backbone's internal priors denoise the boundary noise. Semantic corruptions (Dropout/Flipping) lead to a performance collapse. This indicates that a primary threat to conditional diffusion reliability is not visual noise, but semantic misinformation implied through the conditioning inputs.

3.3.2. Observation 2: The Reliability Tipping Point

The reliability-aware method provides maximum utility at moderate noise levels (0.25–0.50). This reveals a "Confidence Threshold" in generative models. At 50% noise, the input is still recognizable but distracting. By reducing Q, we prevent the model from wasting sampling effort on corrupted regions. This means reliability-aware systems are most useful in real-world scenarios where data is imperfect but not yet completely noisy.

3.3.3. Observation 3: Structural Alignment Paradox

In Table 1, Structural Alignment scores increase as noise increases for the baseline. This is an interesting finding because it implies that a higher alignment score in a noisy environment indicates a failure of the system. It proves the baseline is overfitting to the noise, turning random pixels into random objects. The reliability-estimator method intentionally lowers this alignment to prioritize a coherent image, proving that adhering to a corrupted map reduces image quality.

3.4. Failure Cases

A key limitation is observed at 0% noise (clean data), where a slight performance dip occurs. This shows that the estimator occasionally misidentifies geometry (such as thin chair legs, textured surfaces, or overlapping indoor objects) as corruption. Since the reliability estimator is designed to suppress uncertain conditioning regions, highly detailed but correct segmentation structures may incorrectly receive lower confidence values. As a result, the framework slightly

(a) Boundary Noise (Baseline) (b) Region Dropout (Baseline) (c) Label Flip (Baseline)

(d) Boundary + Reliability (e) Region Dropout + Reliability (f) Label Flip + Reliability

Figure 3. **Qualitative comparison of generation results using Sample 3 from Figure 1**. The top row (samples a, b, c) demonstrate the baseline images (conditioned only on the noisy segmentation maps). The bottom row (samples d, e, f) demonstrate the reliability-adjusted images (conditioned on noisy segmentation maps and the reliability estimator Q value). Each sample, contains 4 square images that show the generated images across noise levels 0.0, 0.25, 0.50, 0.75 from the top left square to bottom right square. With greater noise, the image degrades since the structure of the condition loses its clarity. The reliability-aware model produces more coherent images by relying less on the conditional input and more on the prompt. Boundary Noise primarily introduces mild mask noise (which does not largely impact the model's understanding of the segmentation map), Region Dropout removes object regions, and Label Flip introduces semantic contradictions that cause object misplacement. The reliability-aware model dynamically down-weights unreliable conditioning signals, resulting in more stable object presence under Dropout and Label Flip, although at extreme noise levels (0.75), both models degrade due to the random nature of the conditioning input. The figure illustrates how adaptive reliability modulation improves robustness by preventing the model from strictly following corrupted segmentation structure.

reduces the influence of useful conditioning information even when the segmentation map is fully accurate.

Another failure mode occurs at extreme corruption levels (0.75 noise), where the segmentation map becomes nearly stochastic. Although the reliability-aware method still improves semantic consistency relative to the baseline, the estimator cannot recover structural information that no longer exists in the conditioning input. Under these conditions, the model defaults toward the text prior, causing generated layouts to become more generic. This demonstrates that reliability-aware conditioning can mitigate corrupted guidance, but it cannot reconstruct meaningful structure from fully destroyed semantic inputs.

4. Discussion

The results demonstrate that even a simple, non-learned reliability estimator can improve robustness in conditional diffusion models. By adjusting conditioning strength via the Q-metric, the model reduces over-reliance on corrupted segmentation inputs and instead rebalances generation toward the text prior. This behavior is most evident under Region Dropout and Label Flip noise, where semantic inconsistencies in the conditioning map would otherwise mis-

Noise Type	Level	Avg CLIP (Semantic)	Avg Alignment (Structural)
Boundary	0.00	25.50	0.108
	0.25	25.01	0.108
	0.50	25.19	0.107
	0.75	25.13	0.110
Dropout	0.00	25.54	0.107
	0.25	20.41	0.277
	0.50	19.36	0.296
	0.75	17.30	0.286
Label Flip	0.00	25.08	0.107
	0.25	21.75	0.276
	0.50	18.60	0.297
	0.75	17.67	0.308

Table 1. **Model Robustness across noise types and noise levels.** Semantic consistency is measured using CLIP similarity between the generated image and the text prompt, while Structural Alignment measures adherence to the segmentation condition. Boundary Noise causes little change in semantic performance across all noise levels, suggesting that the diffusion backbone naturally smooths small geometric perturbations. In contrast, Region Dropout and Label Flipping significantly reduce CLIP similarity as noise increases, demonstrating that ControlNet is more sensitive to semantic corruption. Structural Alignment scores increase under severe Dropout and Label Flip noise, indicating that the baseline model increasingly overfits to corrupted conditioning maps rather than generating semantically coherent images. These results show that semantic corruption in conditioning inputs is a major failure mode for conditional diffusion models.

Noise Type	Level	Baseline CLIP	Rel.-Aware CLIP	Improvement (%)
Boundary	0.00	25.50	25.60	+0.4%
	0.25	25.01	26.03	+4.1%
	0.50	25.19	25.73	+2.1%
	0.75	25.13	25.43	+1.2%
Region Dropout	0.00	25.54	25.02	-2.0%
	0.25	20.41	22.80	+11.7%
	0.50	19.36	20.51	+5.9%
	0.75	17.30	18.59	+7.4%
Label Flip	0.00	25.08	25.01	-0.3%
	0.25	21.75	23.31	+7.1%
	0.50	18.60	21.94	+18.0%
	0.75	17.67	18.10	+2.4%

Table 2. **Impact of Reliability-Aware Conditioning on semantic alignment under varying segmentation noise conditions**. The reliability-aware framework adjusts conditioning strength using the estimator Q. Improvements are minimal under Boundary Noise because the baseline model is already robust to mask corruption. Significant gains occur under Region Dropout and Label Flip noise, where semantic inconsistencies degrade baseline performance. The largest improvement occurs at 50% Label Flip noise, where CLIP similarity increases by 18.0%, showing that the framework successfully prevents over-adherence to contradictory labels. Gains diminish at 75% noise because the segmentation map contains little recoverable structure. Slight decreases at 0% noise suggest the estimator occasionally suppresses valid geometric detail by interpreting complex structures as unreliable. Adaptive conditioning is most effective in moderate-noise regimes where the input is degraded but still partially informative.

guide the model.

The implication is that current conditional diffusion models behave as overly compliant systems. They strongly trust the conditioning signal even when it is semantically incorrect. The introduction of a reliability-aware mechanism adds a form of controlled skepticism, allowing the model to down-weight unreliable inputs rather than blindly following them. This shift is important in cases like Label Flip noise, where the structural layout remains coherent but semantically inverted, a failure mode that standard ControlNet consistently overfits to.

This work highlights a limitation of existing ControlNet-

based systems: they lack an internal mechanism for assessing the validity of their conditioning inputs. Without some reliability-estimator, they cannot distinguish between informative structure and misleading or corrupted supervision. Introducing reliability-aware conditioning is a step toward making diffusion models more adaptive and self-correcting, which is key for transitioning from controlled experimental settings to real-world applications such as interactive design tools, autonomous scene editing, and AR-based content generation, where imperfect or noisy inputs are the norm rather than the exception.

4.1. Ethical and Societal Considerations

Diffusion-based generative models have a wide range of applications across design, art, and data synthesis, but they also raise important concerns with misuse and bias. While the reliability-aware conditioning mechanism improves the robustness of generated outputs, it may also be misapplied in contexts involving unanticipated or sensitive data sources, including non-consensual surveillance imagery. In addition, the heuristic nature of the Q-metric introduces potential bias, as complex or culturally specific structures may be interpreted as noise, leading to unintended suppression of structural diversity in the outputs.

Mitigation strategies such as digital watermarking in downstream applications may help improve traceability. At a broader level, this method is intended to support human–AI collaboration in creative domains by improving robustness to imperfect inputs, rather than to operate in sensitive settings without careful bias evaluation and informed consent. Extensive validation is necessary before the model is deployed in contexts where fairness and cultural representation may be affected.

4.2. Conclusion

In this work, the robustness of ControlNet-based diffusion models under noisy segmentation conditioning is explored. The experiments show that segmentation corruption significantly degrades generation quality, which demonstrates a key limitation of current conditional diffusion frameworks.

To address this limitation, a simple reliability-aware conditioning strategy is proposed to estimate segmentation quality using edge density and structural fragmentation, adapting conditioning strength accordingly during inference. The results demonstrate that this approach improves robustness at moderate noise levels by reducing over-reliance on corrupted inputs.

From a zoomed out perspective, this work emphasizes the importance of incorporating input quality awareness into conditional generation pipelines. While improvements are modest and performance remains limited under extreme noise, the proposed method provides a practical step toward more robust generative systems. Future work could explore learned reliability estimators, integration of uncertainty into diffusion training, and more advanced mechanisms for handling severely degraded inputs. Overall, this research moves us closer to deploying more diffusion-based models in real-world applications, such as medical imaging and professional design workflows, ultimately enabling impactful and reliable uses of this technology.

4.3. Use of GenAI

I used ChatGPT to help with brainstorming a reasonable structure of the paper. I also used it to help with some code generation/debugging.

References

[1] Md Manjurul Ahsan, Shivakumar Raman, Yingtao Liu, and Zahed Siddique. A comprehensive survey on diffusion models and their applications, 2024. 2

[2] Shariq Farooq Bhat, Niloy J. Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning, 2023. 2

[3] K Hadiyya, R Dina, and B Mokhtar. Diffusion models in generative ai: Principles, applications, and future directions. *Preprints*, 2025. 1

[4] Moshe Kimhi, Omer Kerem, Eden Grad, Ehud Rivlin, and Chaim Baskin. Noisy annotations in semantic segmentation, 2025. 1, 3

[5] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, 2015. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[8] Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. Dc-controlnet: Decoupling inter- and intra-element conditions in image generation with diffusion models, 2025. 2

[9] Matthew Yee-King. *Latent Spaces: A Creative Approach*, pages 137–154. Springer International Publishing, Cham, 2022. 1

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 3

[11] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 3

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 2018. 3