Risk-Aware Image Generation by Estimating and Propagating Uncertainty

Alejandro Perez [1] **Iaroslav Elistratov** [1] **Fynn Schmitt-Ulms** [1] **Ege Demir** [1] **Sadhana Lolla** [1]
Elaheh Ahmadi [1] **Daniela Rus** [1] **Alexander Amini** [1]

Abstract

While generative AI models have revolutionized content creation across various modalities, they have yet to be deployed in safety-critical scenarios. This is in part due to limited understanding of their underlying uncertainty, as general-purpose frameworks for estimating uncertainty in large-scale generative models are lacking. Here we analyze the effects of uncertainty and risk estimation methods on generative AI systems and their applications to two critical domains of deployment – identification of failures, and fast optimization of input prompts. As a case study, we apply our approach to create an uncertainty-aware variant of the Stable Diffusion text-to-image model, allowing us to estimate and propagate uncertainty over inputs, latent representations, and outputs. We demonstrate that our method enables the identification of uncertain output regions and the optimization of input prompts to minimize output uncertainty. We envision that our framework will enable the deployment of more robust and auditable generative AI systems.

1. Introduction and Related Work

While generative AI has shown impressive capabilities, including in image creation, they are susceptible to certain failure modes such as hallucinations and adversarial attacks. As we consider the deployment of generative AI models in safety-critical scenarios and society at large, it becomes crucial to develop models that can reliably assess their risks and uncertainties as well as leverage this understanding to improve robustness during deployment. This notion of risk awareness is necessary to ensure the robustness, understand the limitations, and maximize the quality of generative AI models.

[1]Themis AI. Correspondence to: Alexander Amini <amini@themisai.io>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

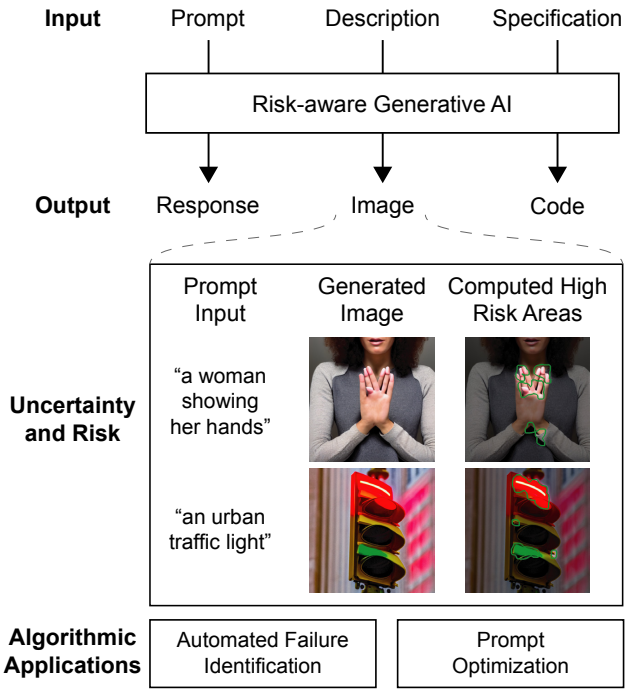

Figure 1. **Risk-aware image generation.** By endowing generative models with uncertainty-aware capabilities, we demonstrate automated (1) identification of highly uncertain output regions and (2) optimization of input prompts to minimize output uncertainty, thereby maximizing output quality. Our results provide principled insights into challenging aspects in generative AI deployment.

Existing approaches for estimating uncertainty in AI models estimate a singular form of uncertainty in the context of specific data modalities (Nix & Weigend, 1994; Kendall & Gal, 2017; Lakshminarayanan et al., 2017; Buolamwini & Gebru, 2018; Zhang et al., 2018a).

Here we present an uncertainty-aware framework that enhances generative AI systems for risk-assessment in deployment scenarios. Our method enables the identification of highly uncertain output regions as well as the optimization of input prompts to minimize output uncertainty. As a case study, we apply this approach to create a risk-aware variant of the Stable Diffusion text-to-image model. Our method allows us to generate and propagate uncertainty estimates

over inputs, outputs, and latent representations, providing valuable insights into the uncertainties associated with the generative process.

In summary, our work makes the following contributions:

1. A scalable method for estimating and propagating uncertainty through large generative AI models to uncover and diagnose various issues that may arise during deployment.
2. A risk-aware variant of the Stable Diffusion model that is capable of estimating and propagating uncertainty through its text inputs, latent encodings, and image outputs.
3. An analysis of the uncertainties associated with inputs (text) and outputs (generated images), shedding light on the challenges and opportunities in deploying risk-aware generative systems.
4. A prompt optimization algorithm that improves the output image quality by iteratively reducing uncertainty in the input prompts.

2. Experimental Setup

2.1. Stable Diffusion Model

We consider a Keras implementation of Stable Diffusion (Chollet et al., 2015; Rombach et al., 2021; 2022; Gupta, 2022), a latent text-to-image diffusion model. The image generator contains three main models which we individually estimate and propagate uncertainty through, i.e., CLIP (Radford et al., 2021), a text encoder that takes in prompts as input, the diffusion model (Ho et al., 2020), used to generate latents over multiple steps, and an autoencoder decoder (Rombach et al., 2021), which converts latents into images. We provide a brief outline of the architecture below.

The text descriptions provided by the user, i.e., prompts, are processed by CLIP (Radford et al., 2021; Ilharco et al., 2021). This text encoder produces embedding vectors in a 768-dimensional space for each token in the prompt. The diffusion model (Ho et al., 2020) takes in these token embeddings, i.e., text description of what will be in the image, and processes that information in a latent space iteratively over 50 steps, starting from random noise, in a reverse diffusion process. The latent diffusion model is implemented as a U-Net (Ronneberger et al., 2015). At each step the model outputs $[64, 64, 4]$ latents. After the diffusion process, the resulting latent is passed into the image decoder to generate the final image. An autoencoder decoder (Esser et al., 2021; Zhang et al., 2018b; Yu et al., 2021) with several ResNet (He et al., 2016), up-sampling, and convolutional layers take in latents to produce the final $[512, 512]$ images. Decoding is the final step of the image generation process and takes place only once in the current implementation. For more information on the details of the Stable Diffusion model we refer readers to the original publication (Rombach et al., 2021).

2.2. Risk and Uncertainty Estimation

Bayesian neural networks (Blundell et al., 2015) and other epistemic uncertainty estimation methods (Lakshminarayanan et al., 2017; Kendall & Gal, 2017; Gal & Ghahramani, 2016; Amini et al., 2020) enable neural networks to algorithmically estimate their own predictive uncertainty. Many of such methods operate by sampling from a probabilistic distribution of weights with each sample producing a unique model instantiation – and thus a unique answer for a fixed input. By evaluating the variability across sampled outputs, we can estimate the predictive uncertainty of the model for a given input. Such sampling operations can be performed across any scale of the model, making each of the three components (CLIP, latent diffusion, and image decoding) entirely risk-aware.

Traditionally, models output predictions in the form of $\hat{y} = f_{\boldsymbol{W}}(x)$. By applying a risk-aware transformation, Φ, we build a risk-aware variant, such that

$$g = \Phi_{\boldsymbol{\theta}}(f_{\boldsymbol{W}}),$$
$$\hat{y}, R = g(x),$$

where R are the estimated risk measures from a set of metrics, $\boldsymbol{\theta}$.

We then take the resulting risk-aware model and proceed to generate novel samples from it. However, equipped with risk-awareness, our models will now be able to additionally estimate uncertainty for every output that it predicts. Across these generations we compute uncertainties and store them accordingly. We assess results and performance by evaluating the inputs text in the context of the encoded uncertainties as well as the uncertainties that arise during the iterative latent decoding process.

We use the Capsa framework (Lolla et al., 2022) to perform model transformations. We refer readers to the open-source version of the Capsa library (Amini et al., 2022) for an introductory description of some procedures and to Capsa Pro (Amini et al., 2023) for information on the software library with the functionality described in this publication.

2.3. Risk-Aware Prompt Optimization

Writing natural language prompts that produce desired results can be a difficult process that often requires the user to go through tedious trial-and-error experimentation or to consult resources written by experts (Zamfirescu-Pereira et al., 2023). In an effort to address this issue, several approaches for automated prompt engineering have been

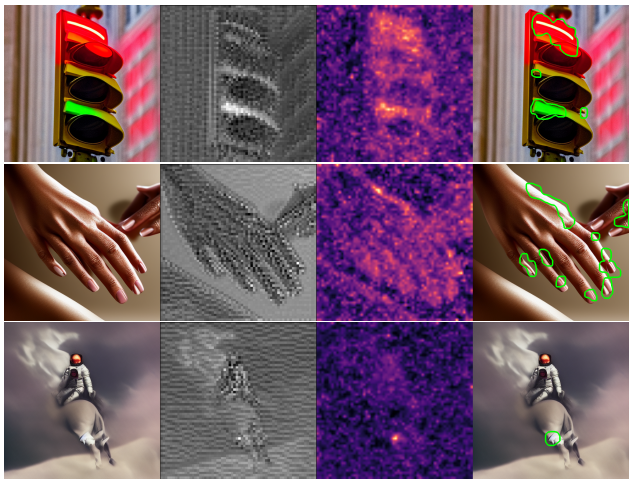

Figure 2. **High-uncertainty regions from outputs of risk-aware Stable Diffusion.** Left-to-right: generated image, latent diffusion state, estimated pixel-wise uncertainty, highest-uncertainty threshold over the image. In column 3, darker and lighter are lower and higher uncertainty respectively.

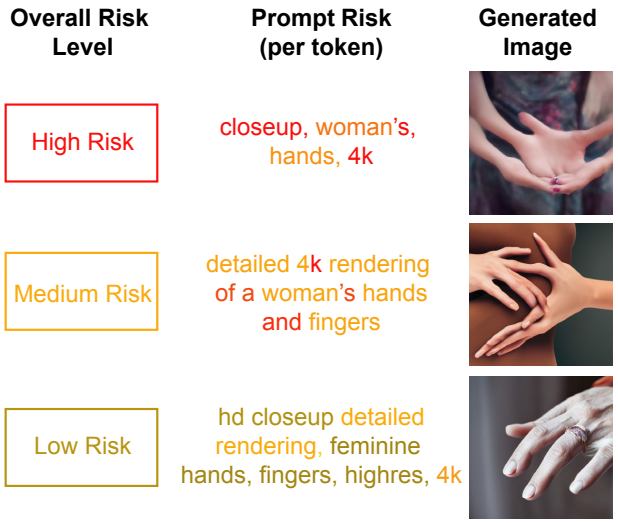

Figure 3. **Assessment of token-by-token uncertainty in text prompts for image generation.** Individual prompts are evaluated for overall uncertainty (left) and per-token uncertainty (middle). Varying levels of uncertainty result in varying levels of fidelity in the generated images (right).

proposed (Pryzant et al., 2023; Zhou et al., 2022; Zhang et al., 2023). However, these methods rely on additional training, require access to either training data or internal state variables, or utilize separate language models in the process.

Because we have converted our text encoder into a risk-aware variant, we are able to utilize a simple algorithm that considers only the output uncertainty of the text encoder itself to optimize any prompt. The approach we used is as follows.

30 brief descriptions of a desired output image are used as the base text for prompts. We collect several prompt examples from an open-source prompt engineering database and identified the 50 most common phrases used to increase the quality of generated images. We then define a list of operations the algorithm is able to perform on a given prompt. These are: sample and remove one word, change the location of a word within a prompt, replace a word with a synonym, and add a word from the base text. We use our uncertainty-aware text encoder to evaluate each prompt, using its output uncertainty as the metric. The procedure iterates as a Monte Carlo Tree Search (Coulom, 2007) algorithm with branch-and-bound (Kumar et al., 1988). That is, each prompt is a state in the tree. During each iteration, a state is sampled along with one operation used to modify it. We then evaluate the output uncertainty after the modification and add new states to the tree only if they result in lower uncertainty estimates. The results considered in this publication were obtained by using this procedure to generate 10,000 prompts.

Using the approach described above, we can see that the algorithm is able to combine the words in the base text descriptions to form prompts that lead to lower uncertainty estimates for the text encoder, resemble prompts created by prompt engineers, and lead to realistic, high-quality images depicting what is described in the text.

3. Results

3.1. Uncertainty in Image Diffusion Models

We begin our empirical analysis by prompting and sampling from our risk-aware Stable Diffusion model, and in turn evaluating the outputs and the associated uncertainties (Figure 2). Specifically, we measure the pixel-wise uncertainty across the inferred latents and the decoded output, and use computed pixel-wise scores to estimate thresholds of highest uncertainty over the generated image. In Figure 2, we show results from three representative examples.

Across several classically challenging prompts (e.g., stoplights, human hands (Borji, 2023)), we observe that regions estimated to have high uncertainty correspond to semantically challenging, fallacious, or incorrect regions of the generated image (Figure 2). In the stoplight example, high uncertainty scores are found in the red and green colored regions, which are completely misplaced above the bulb of the stoplight. In the human hand example – a known failure mode for the most powerful generative image models – we observe targeted regions of high uncertainty around the nails and extra fingers in the generated image. Finally, in the example of an astronaut riding a horse, the method returns

Figure 4. **Uncertainty-guided prompt optimization improves image generation.** Top: images generated using the original prompt "`an ostrich by a lake`". Bottom: images generated after optimizing input prompts to minimize their own uncertainty. By optimizing prompts to minimize their uncertainty, we demonstrate that downstream generated image quality is also improved as a result – without the need to manually perform hand-engineered prompt tuning and engineering.

a single focal region of high uncertainty, which upon close inspection corresponds to a region of spacesuit misplaced on the horse's leg. Together, these results highlight the ability of the risk-aware Stable Diffusion model to identify semantically meaningful regions of high uncertainty in generated output images.

3.2. Risk-Aware Prompt Optimization and Generation

Having shown that our risk-aware diffusion model enables direct interpretation of uncertainty and risk in generated images, we next explored how these uncertainties related to the text-based input prompt. Because our method propagates uncertainty throughout the generative process, we investigated whether our algorithm could also be used to understand uncertainties over the prompt itself, and whether this knowledge in turn could be used to optimize the prompts to produce more predictable image output.

Using our uncertainty-aware text encoding module, we calculated the uncertainties over individual tokens, and visualized the relative uncertainty across prompts (Figure 3). Using the generation of human hands as a case study, we observed that prompts with greater overall and per-token uncertainty resulted in poorer generations (Figure 3). Indeed, empirically, more specific components to a prompt (e.g., "closeup", "feminine hands", "highres") improved the fidelity of the generated image.

This relationship between per-token uncertainty and the quality of the image output led us to hypothesize that uncertainty-guided prompt tuning could improve the model's generations. We devised an algorithm that leverages estimated per-token uncertainty for prompt optimization. We deployed our algorithm over a series of sample prompts in order to optimize them. At regular intervals in the optimization process we leveraged the resulting prompts as input to the risk-aware Stable Diffusion model, generated samples, and then evaluated the quality of image generations as a function of prompt tuning. In a representative example related to the base prompt "an ostrich by a lake" (Figure 4), we observe that the outputs resulting from a uncertainty-optimized prompt are of high fidelity and quality. Our uncertainty-guided prompt optimization led to a significant improvement in the quality of generations relative to the naive, unoptimized base prompt (Figure 4). Taken together, these results demonstrate that the estimated uncertainties from our risk-aware generative model can be deployed successfully for in-the-loop, uncertainty-guided prompt optimization, yielding image generations of greater robustness and fidelity.

4. Conclusion

Here we present an approach for risk-aware image generation by estimating and propagating uncertainty through the Stable Diffusion model. Our approach enables propagation of uncertainty estimates throughout the diffusion-based generative process. We create an uncertainty-aware variant of Stable Diffusion across its CLIP-based encoding, latent diffusion, and image decoding components. In evaluation we observe higher uncertainties in regions of generated images associated with challenging, fallacious, or incorrect content. We develop an algorithm for uncertainty-guided prompt optimization, and demonstrate that token-by-token uncertainties can be used to iteratively refine prompts to ultimately improve the quality of downstream generations.

Promising avenues for further exploration include expanding and applying our approach to other large-scale generative AI models, such as large language models, and incorporating abilities to estimate additional risk metrics. We also aim to deploy our methods in safety-critical scenarios

where in-the-loop risk estimation could help establish the robustness of large-scale generative AI systems. We envision that these results will inspire further development of robust, risk-aware generative AI.

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

Amini, A. et al. Capsa, 2022. URL https://github.com/themis-ai/capsa.

Amini, A. et al. Capsa pro, 2023. URL https://themisai.io/capsa-pro.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Borji, A. Qualitative failures of image generation models and their application in detecting deepfakes. *arXiv preprint arXiv:2304.06470*, 2023.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Chollet, F. et al. Keras, 2015. URL https://github.com/fchollet/keras.

Coulom, R. Efficient selectivity and backup operators in monte-carlo tree search. In *Computers and Games: 5th International Conference, CG 2006, Turin, Italy, May 29-31, 2006. Revised Papers 5*, pp. 72–83. Springer, 2007.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gupta, D. Stable diffusion in keras, 2022. URL https://github.com/divamgupta/stable-diffusion-tensorflow.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., et al. Openclip. *DOI: https://doi.org/10.5281/zenodo*, 5143773, 2021.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kumar, V., Nau, D. S., and Kanal, L. N. A general branch-and-bound formulation for and/or graph and game tree search. *Search in Artificial Intelligence*, pp. 91–130, 1988.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lolla, S., Elistratov, I., Perez, A., Ahmadi, E., Rus, D., and Amini, A. Capsa: A Unified Framework for Quantifying Risk in Deep Neural Networks. In *5th Robot Learning Workshop: Trustworthy Robotics*, 2022. URL https://slideslive.com/38994172.

Nix, D. A. and Weigend, A. S. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Pryzant, R., Iter, D., Li, J., Lee, Y. T., Zhu, C., and Zeng, M. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Rombach, R. et al. Stable diffusion, 2022. URL https://github.com/Stability-AI/stablediffusion.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., and Yang, Q. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2023.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018a.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018b.

Zhang, T., Wang, X., Zhou, D., Schuurmans, D., and Gonzalez, J. E. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.