Generative Concept Security in Trustworthy AIGC

Kun Xu

College of Computer Science and Technology

Nanjing University of Aeronautics and Astronautics

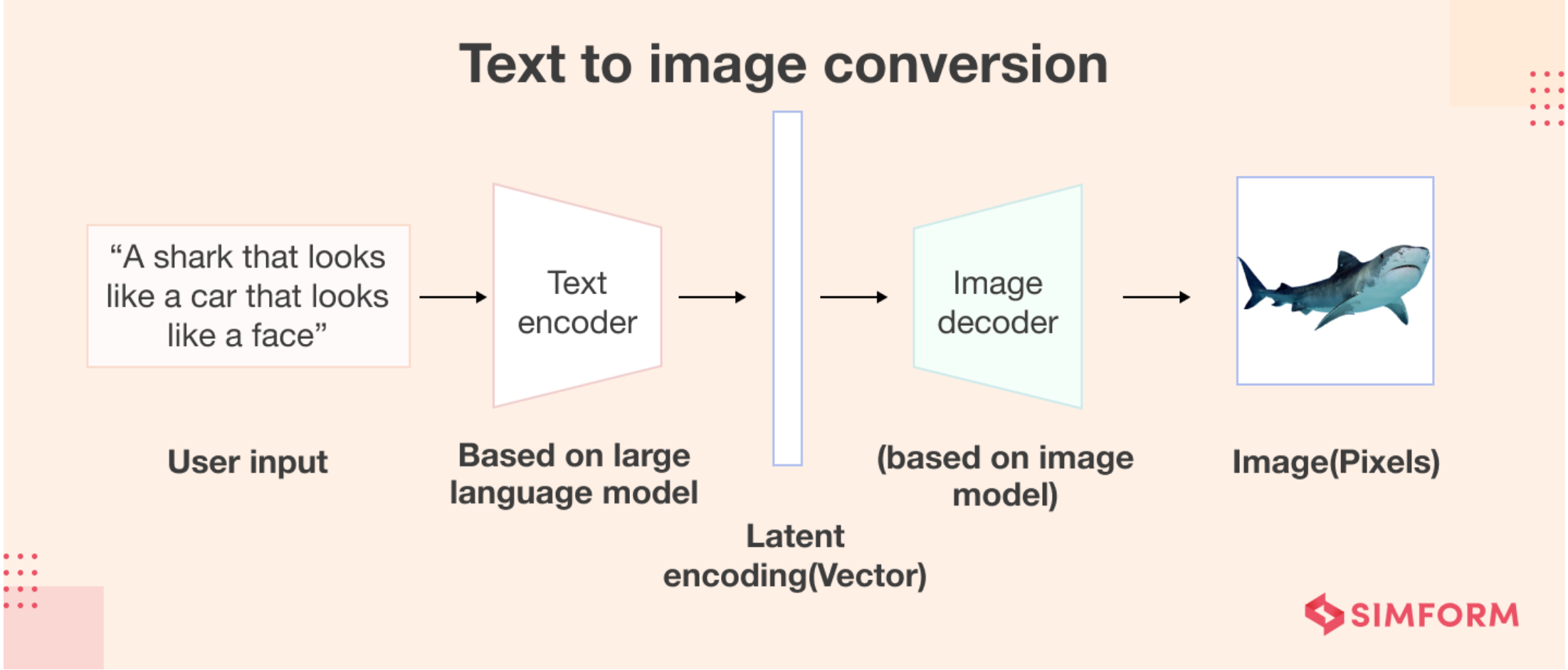
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

The Era of AIGC

This is an AIGC (Artificial Intelligence Generated Content) era. There are many generative models now. Many popular applications.

Text-to-Image (T2I)

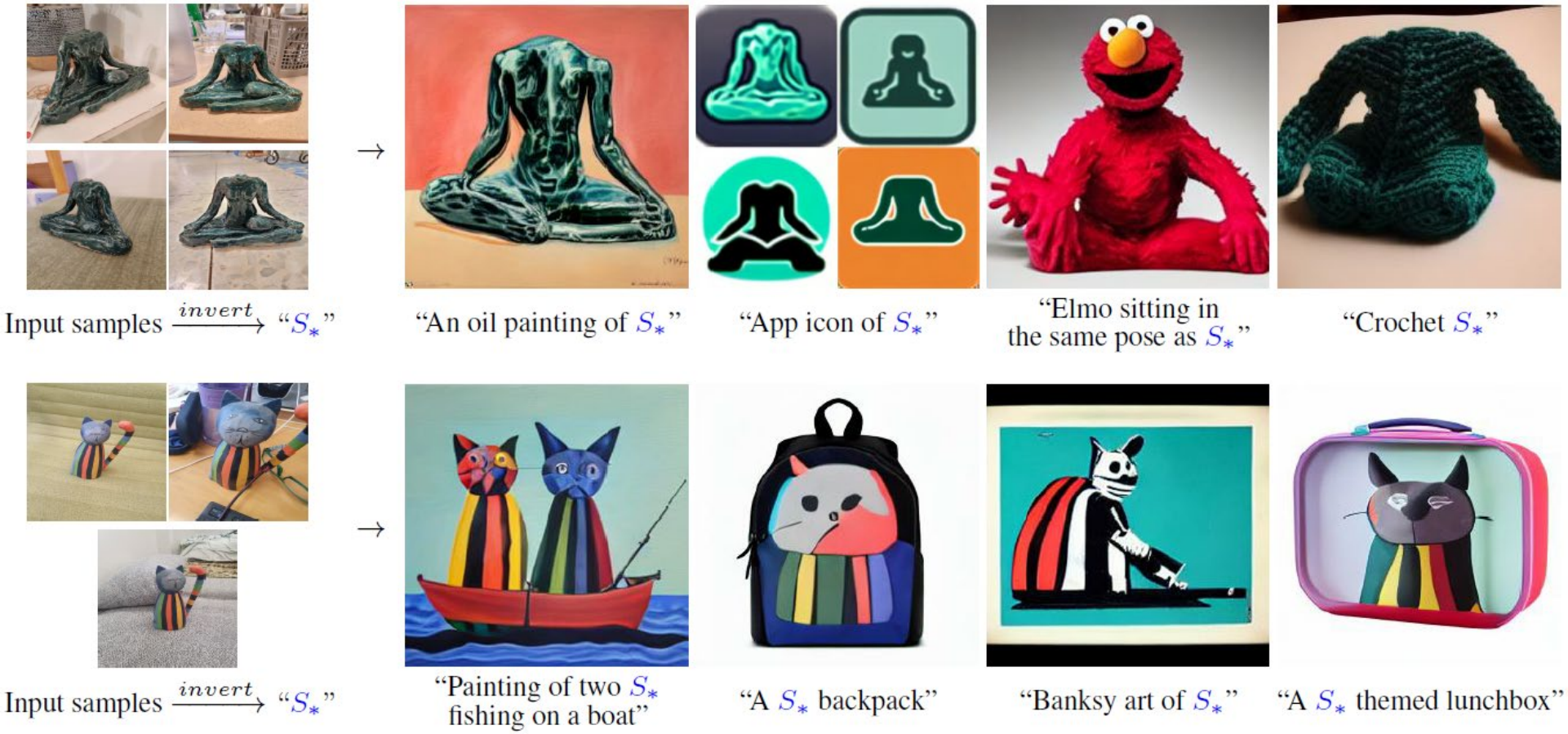
T2I systems interpret natural language prompts and generate corresponding visual content.



Personalized Text-to-Image ⟶ Concept

Concept is considered as a personalized T2I

Textual Inversion



Gal R, Alaluf Y, Atzmon Y, et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. The Eleventh International Conference on Learning Representations.

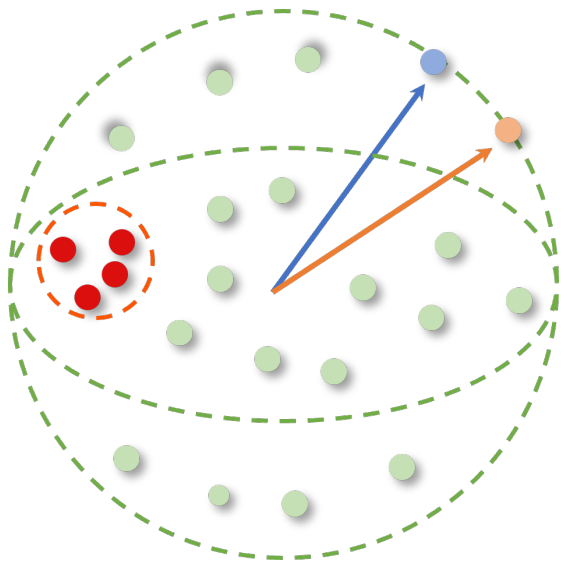
Personalized Text-to-Image ⟶ Concept

Concept is considered as a personalized T2I

DreamBooth

Ruiz N, Li Y, Jampani V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22500-22510.

Personalized Text-to-Image ⟶ Concept



Embedding Space

What is the concept?

- The summary and abstraction of the essential attributes of things is the basic unit of people's cognition of things.
- A mode of thinking that reflects the unique attributes (inherent attributes or essential attributes) of things.

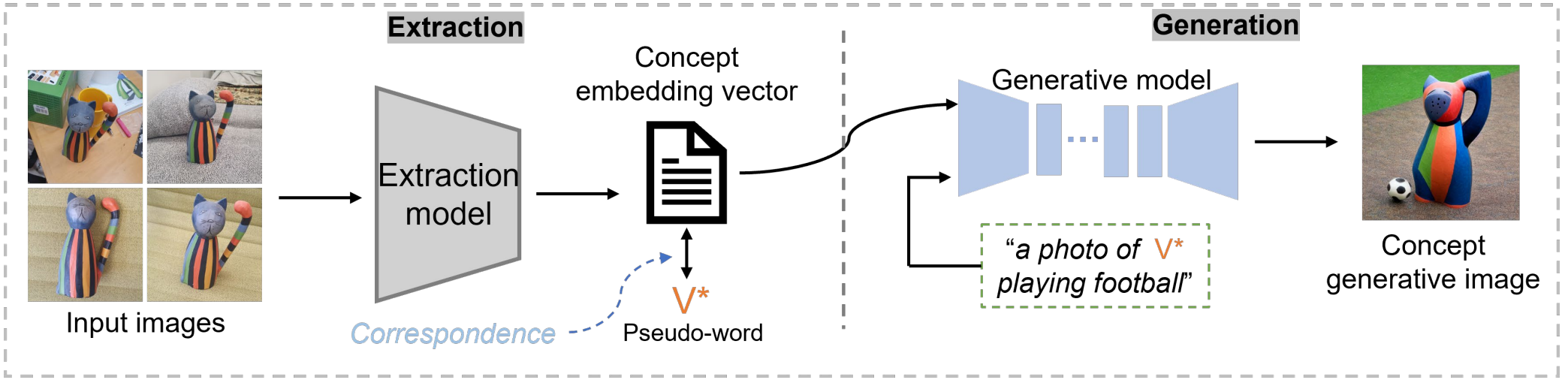
What are concept-driven generative models?

- "Concepts" are explicitly introduced as high-level control and interpretation units in the generation process, making the generation more precise, flexible and interpretable.

Why do image generation models need concepts?

- Limitations of natural language: Natural language cannot accurately describe everything.
- Concepts serve as a supplement to natural language in image generation models and can provide high-level and precise control.

Personalized Text-to-Image ⟶ Concept

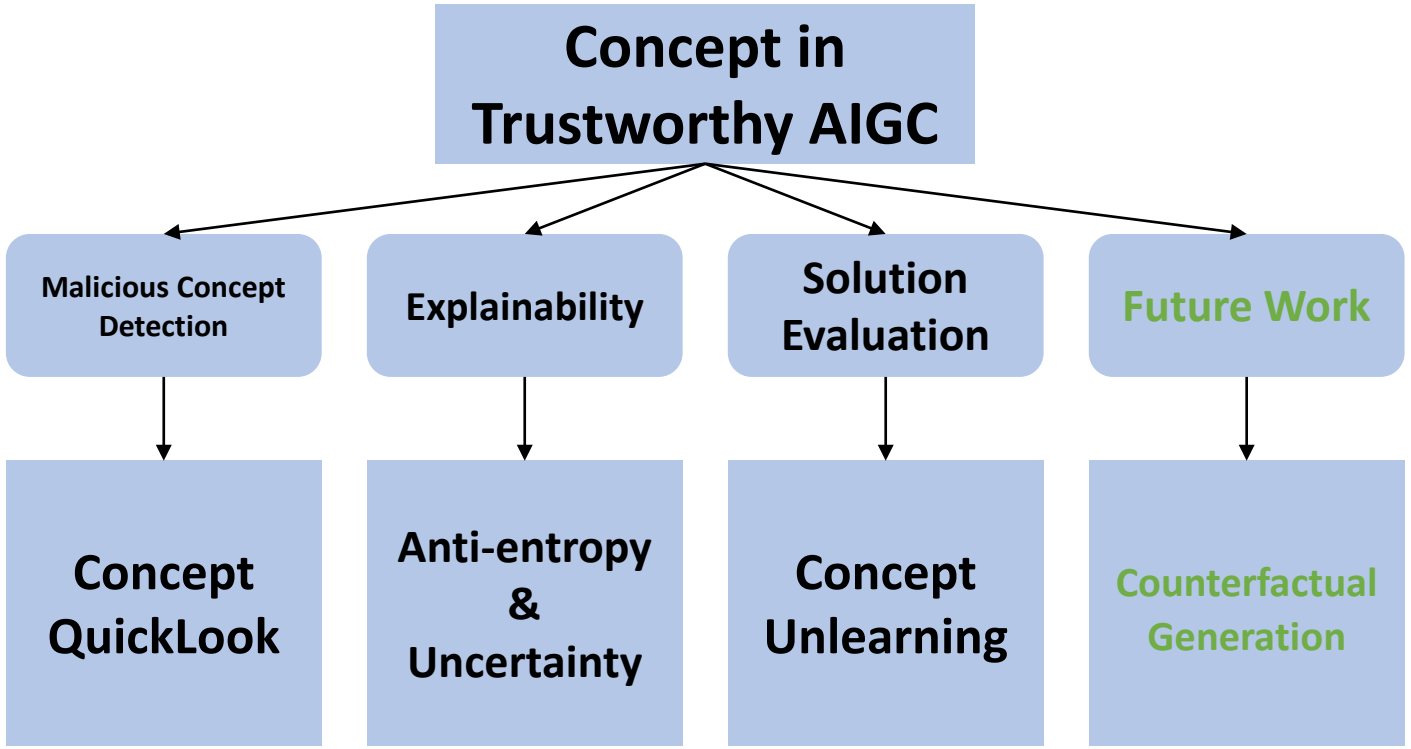


- Limitations of natural language: Language cannot describe everything
- Concept as a supplement to natural language: Concepts can be reproduced in T2I
- There are two processes: Concept extraction and concept generation
- The information representing a concept is stored in the concept embedding vector

My Research Overview

Trustworthy AIGC as **background**

Concept is a **summary** and **abstraction** of things



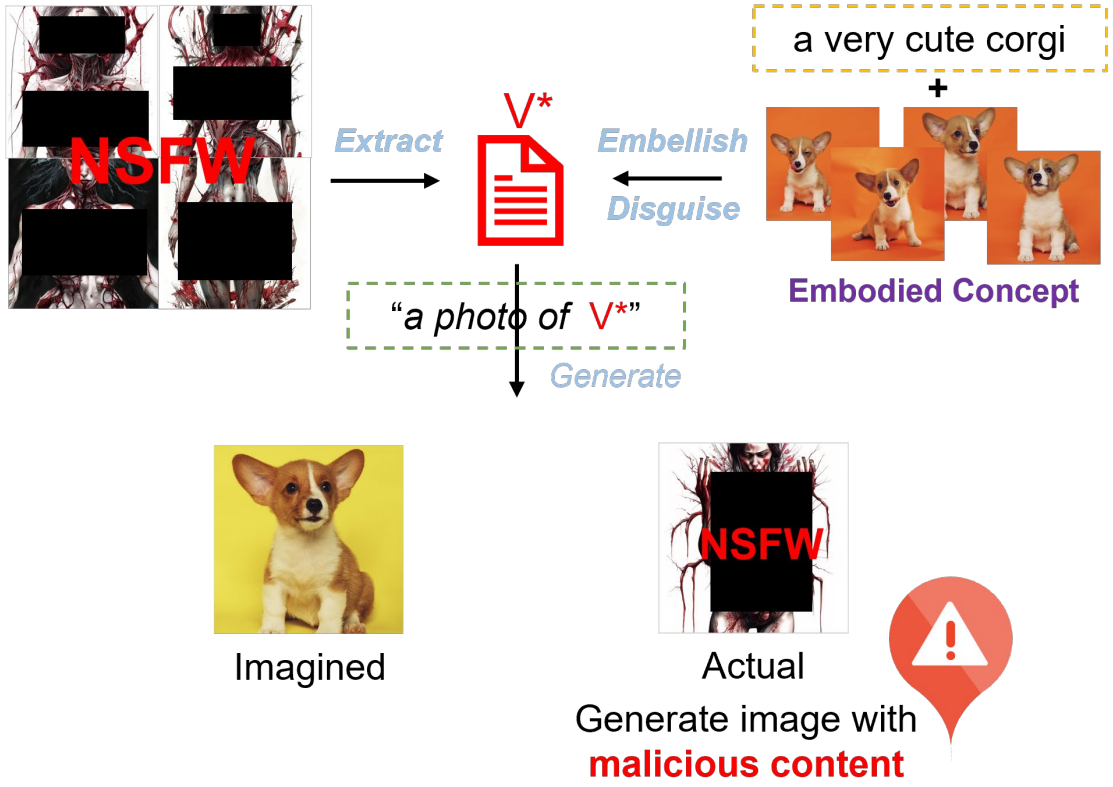
I planned four parts for generative concept research. Each has its own corresponding work, which will be introduced later.

The Work 1

Malicious Concept Detection

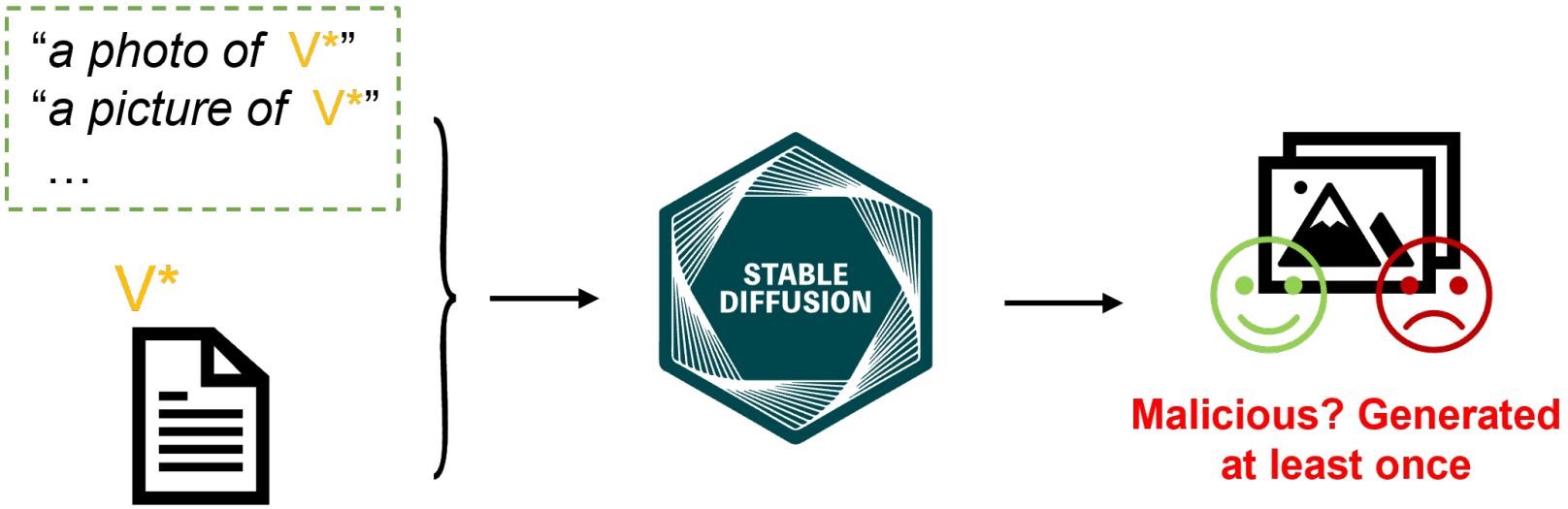
The Risk

Malicious concepts include violence, blood, pornography, etc.



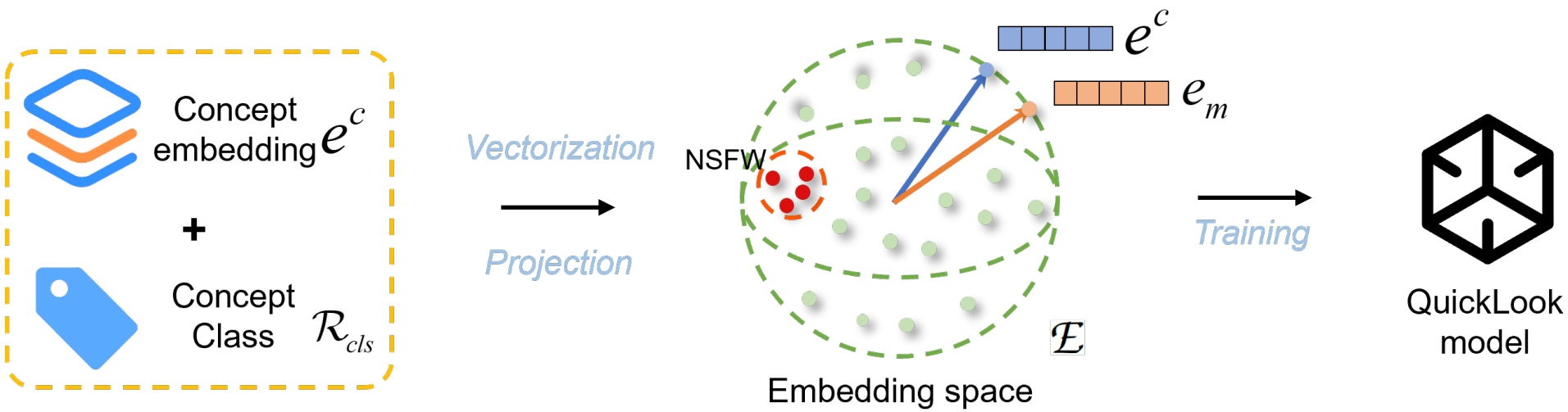
- Concept embedding vectors are non-visual
- Malicious concept: Concept embedding vectors are extracted from the NSFW input images
- Understanding concept embeddings with text descriptions and example graphs
- This description and concept embedding vector relationship is fragile and there is a risk
- Malicious concept embedding vectors are embellished and disguised as normal ones

The Dilemma



- Generating an image at least once can determine whether it is malicious
- Concept generation image judgment has the problem of generating malicious contents
- Inefficiencies and risks

Concept QuickLook

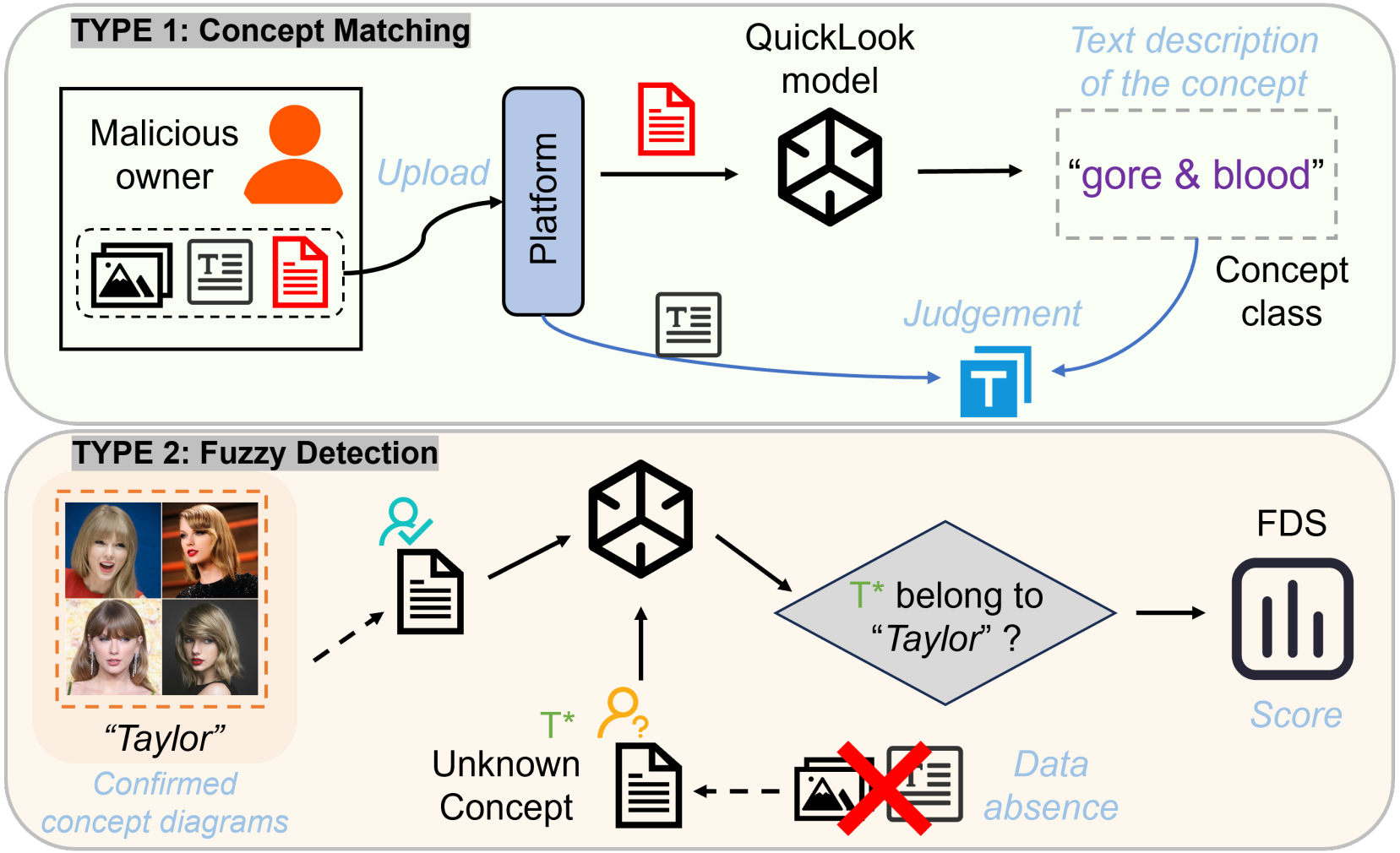


- Extract concept vector
- Encoding concept class

Search embedding space and find the vector that minimizes distance

- The concept of NSFW is also in the embedding vector space

Work Type



Detection consistency with claimed concepts

Detection matches with confirmed concept class

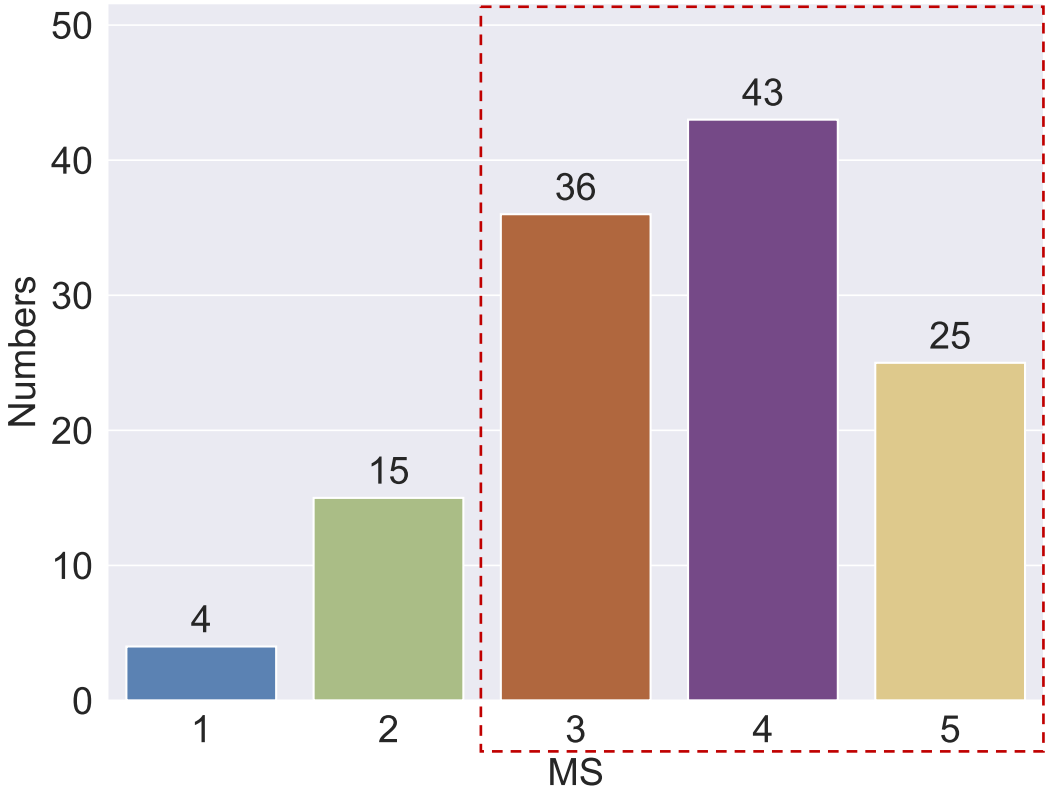
Work Summary

- This work first defines malicious concepts in the concept sharing process and proposes a solution, called Concept QuickLook, to rapidly detect malicious concepts.
- The work analyzes the generation mechanism of the concept generation model and the entire concept file sharing process. It finds that the embedding vectors in the concept files are the primary factor controlling the generated topic content and can be used to detect whether the personalized generated content is malicious.
- Two operating modes are designed for the QuickLook model: concept matching and fuzzy detection. These two modes are demonstrated to effectively meet the requirements for malicious concept detection in current concept sharing platform scenarios.
- Extensive experiments are conducted, including effectiveness evaluation, baseline comparison, manual scoring, and robustness testing. The results demonstrate that the proposed method can identify malicious concepts without requiring a single generation step, effectively protecting the security of concept sharing platforms and their users.

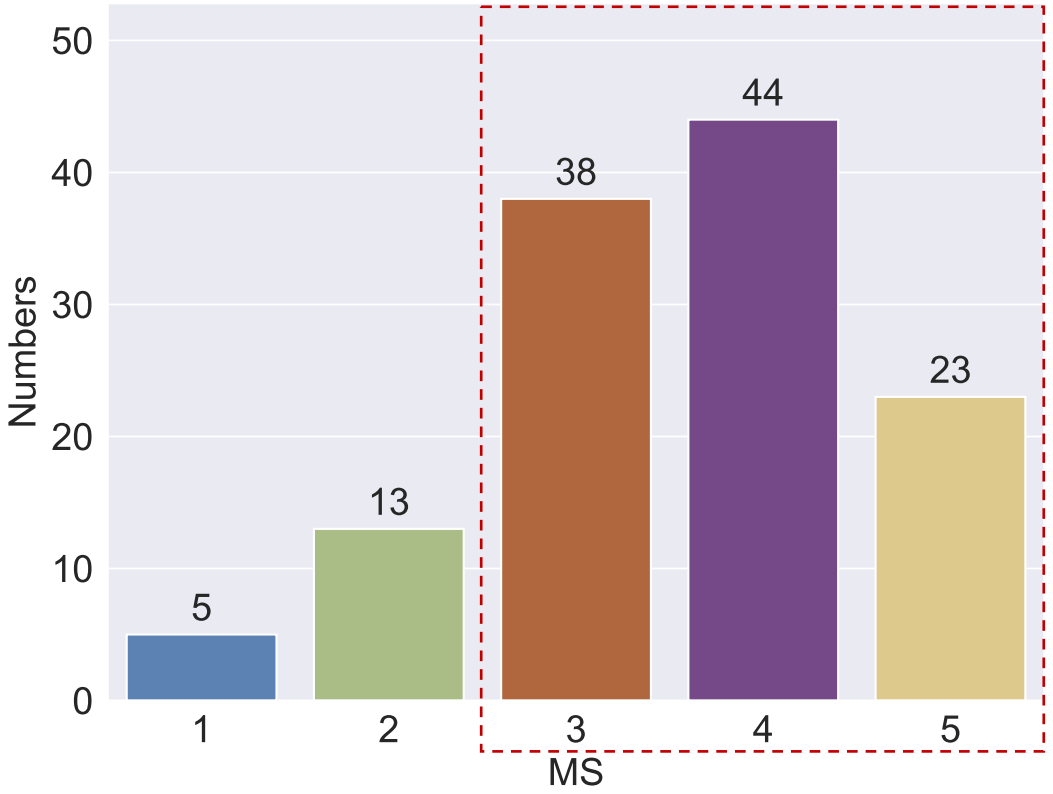
Detection Results

Visual detection results of Concept Matching

Detection Results



Malicious concepts



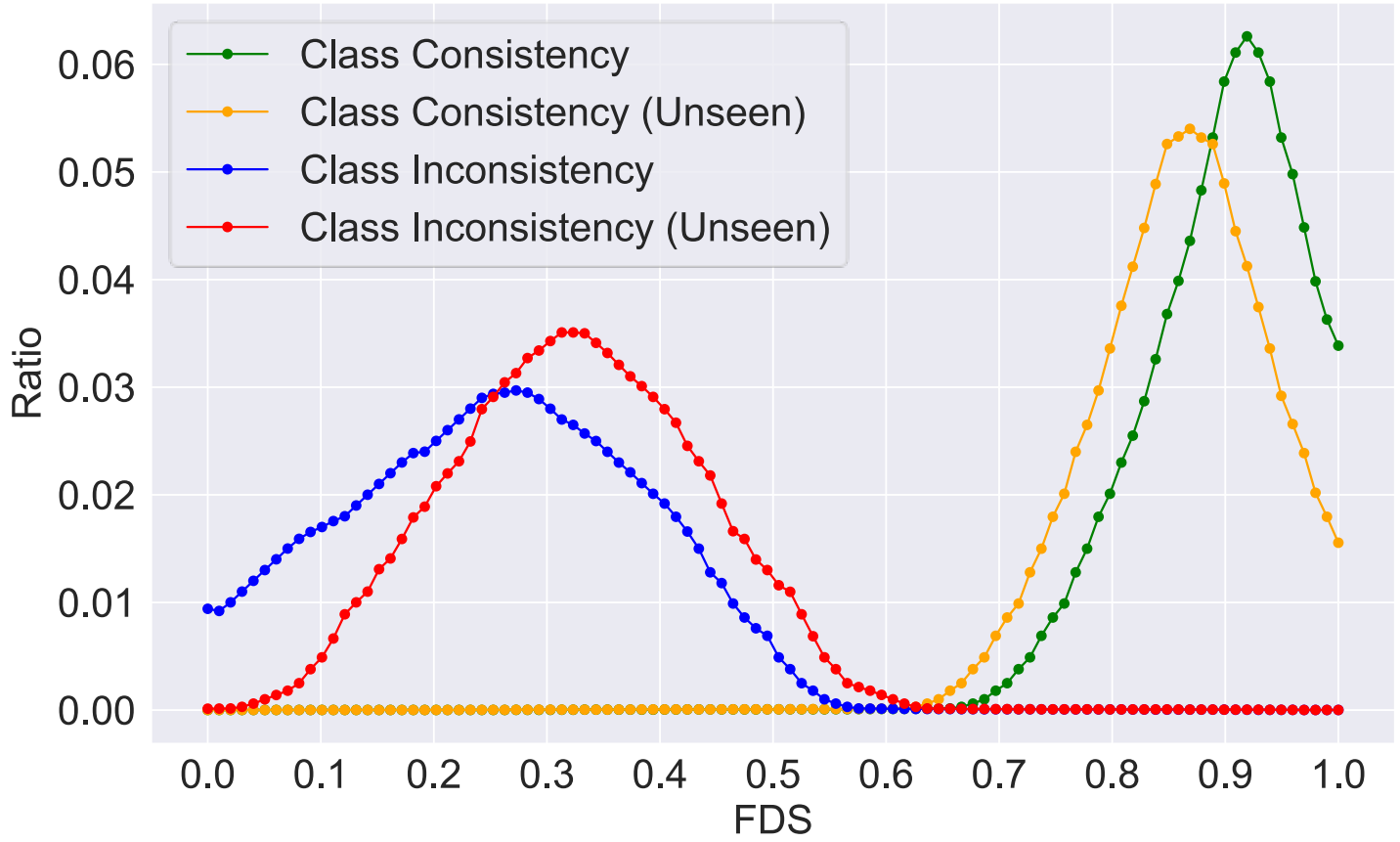
non-malicious concepts

Statistical distribution results of Concept Matching

Detection Results

Visual detection results of Fuzzy Detection

Detection Results



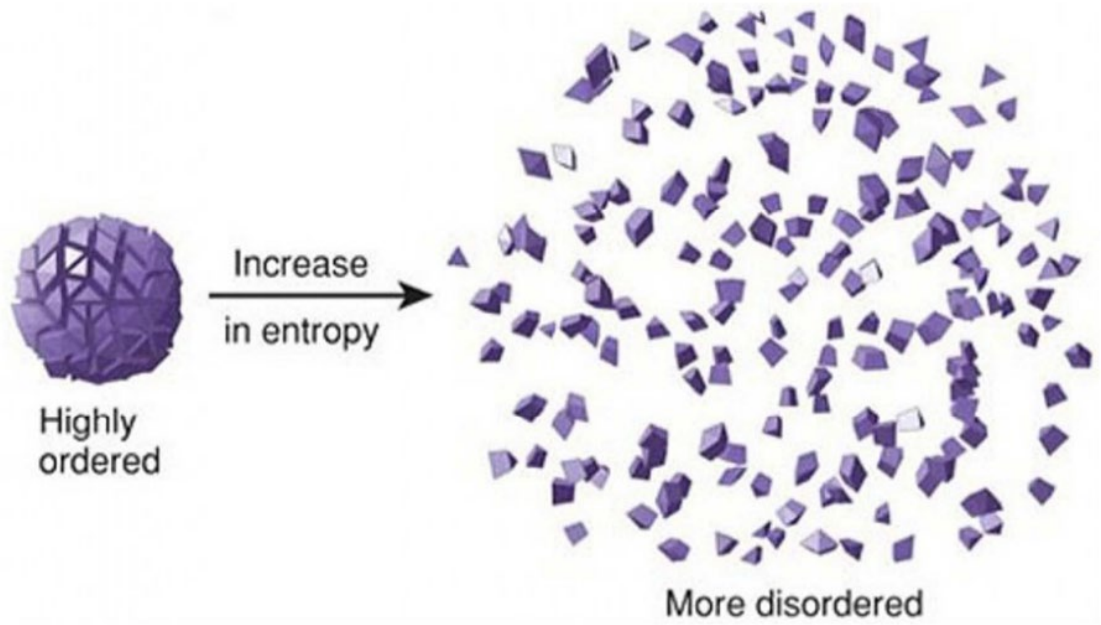
Statistical distribution results of Fuzzy Detection

The peak scores on both sides indicate the highest proportion of consistent and inconsistent concept classes

The Work 2

Understanding Concept-Driven Diffusion Model with Uncertainty

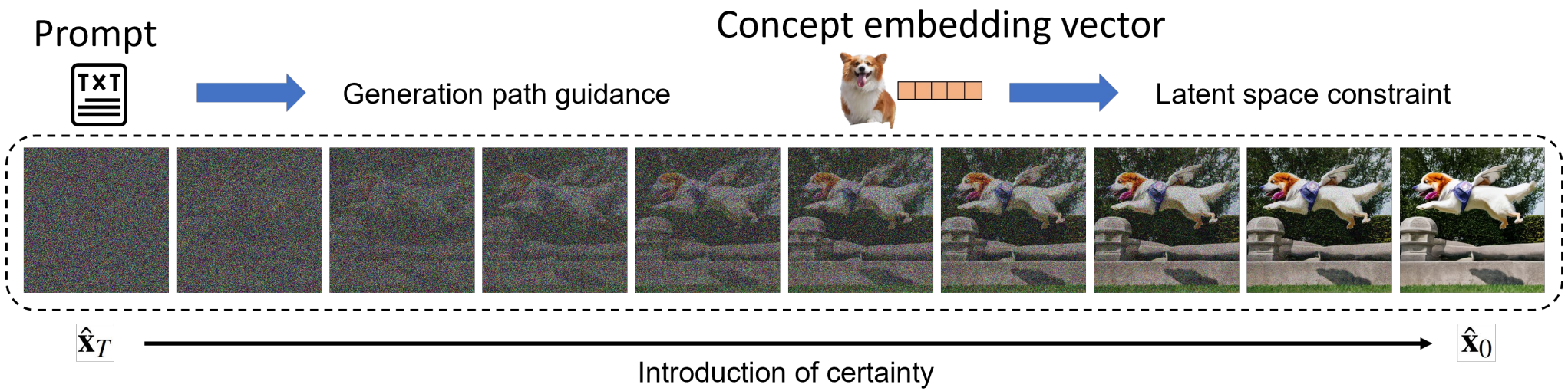
Entropy



The process of entropy increase

- The process of entropy increase is from order to disorder
- The anti-entropy process is from disorder to order
- Certainty flows with entropy
- The forward process of the diffusion model increases entropy, while the reverse process decreases entropy

Understanding Concept-Driven Diffusion Model with Uncertainty



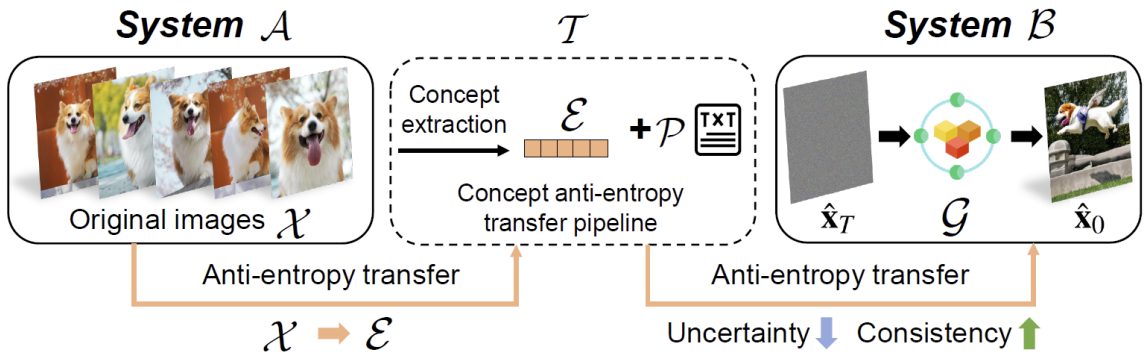
- Concept image generation is a process of certainty introduction
- Concept embedding vector restrict the search space within the latent space to regions near the target concept
- Prompt explicitly indicate the generation direction of the model through textual information

Understanding Concept-Driven Diffusion Model with Uncertainty

Concept Anti-entropy

- Concepts are transferred in the form of embedded vectors, which control the features and direction of the target distribution during the generation process.
- Through progressive denoising, new data representing the target concept are generated from pure random noise.
- This process reduces uncertainty and enhances consistency, reflecting the transfer of anti-entropy from the original images to the embedding vectors and ultimately to the concept generated images.

Understanding Concept-Driven Diffusion Model with Uncertainty



Concept application cycle (CAC)

Concept extraction, concept transfer, concept generation

Hypothesis 1. *The entire CAC from concept extraction through concept transfer to concept generation is a process of anti-entropy transfer* (Sec. IV-A for Hypo. 1).

Hypothesis 2. *Concept generation uses embedding vectors and prompts to introduce certainty into the generation process, thereby reducing the uncertainty of generation* (Sec. IV-B for Hypo. 2).

Concept Unlearning

Unlearning refers to actively removing the influence of certain specific data or knowledge from a trained model so that the model no longer relies on this data or knowledge in subsequent tasks.

- Treat private data as a concept to remove biased or discriminatory knowledge in generative models
- Compliance requirements for Trustworthy AIGC
- Complying with the right to be forgotten

Application:

- Copyright issues involving unauthorized data for AI training (e.g., artist works)
- Evaluation: Verifiable concept unlearning

Concept Unlearning

Unlearning refers to actively removing the influence of certain specific data or knowledge from a trained model so that the model no longer relies on this data or knowledge in subsequent tasks.

Zhang G, Wang K, Xu X, et al. Forget-me-not: Learning to forget in text-to-image diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 1755-1764.

Work Summary

➢ Theoretical Framework of the Concept-Driven Diffusion Model

- Introducing an uncertainty perspective, this work constructs a theoretical framework based on anti-entropy to systematically model the extraction, transfer, and generation of concepts.

- Two key hypotheses are proposed and theoretically verified: the anti-entropy transfer process of concepts and the deterministic introduction of concept generation.

- This framework reveals the underlying mechanisms of concept information flow and uncertainty evolution, providing a novel theoretical perspective for understanding the concept-driven diffusion model.

Work Summary

➢ Concept Uncertainty Quantification Method Based on Anti-Entropy

- This work proposes a unified approach to quantifying semantic and structural uncertainty, supporting multi-granular analysis at both the representational and cue levels.

- This framework can serve as a quantitative tool for assessing the stability, controllability, and generative behavior of concepts in concept-driven diffusion models.

Work Summary

➢ Application of Frameworks and Methods in Concept Unlearning

- This work applies the proposed theoretical framework and uncertainty quantification method to the concept unlearning task, designing and implementing a comprehensive experimental pipeline.
- Through extensive evaluations across different concept representations, generation settings, and unlearning strategies, the proposed framework and method demonstrate their adaptability and practical value.
- This work provides empirical support for evaluating the safety and controllability of generative models.

Results

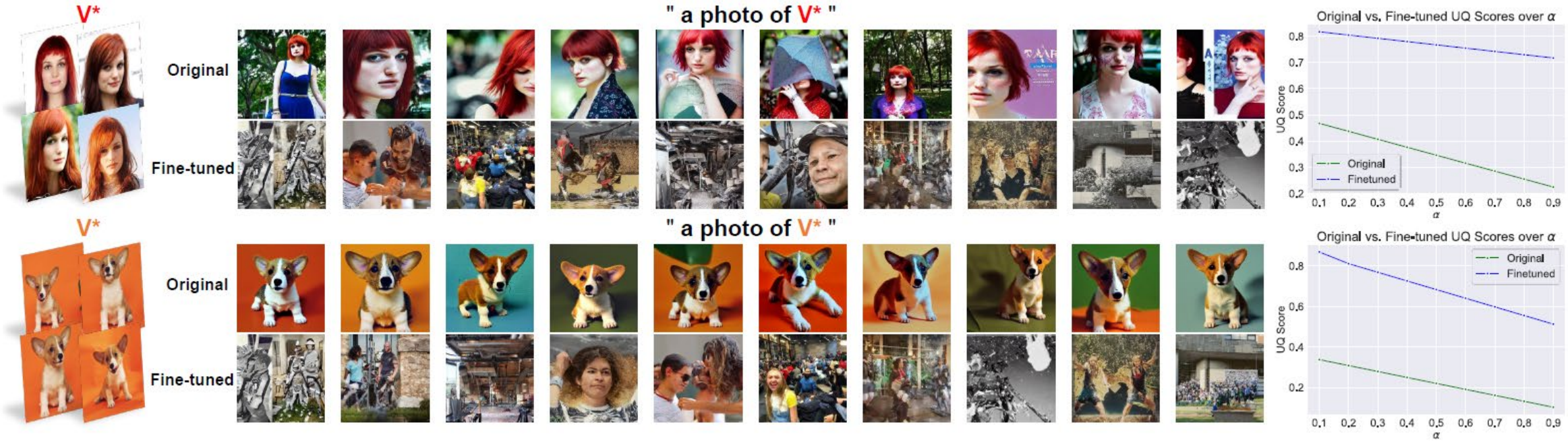
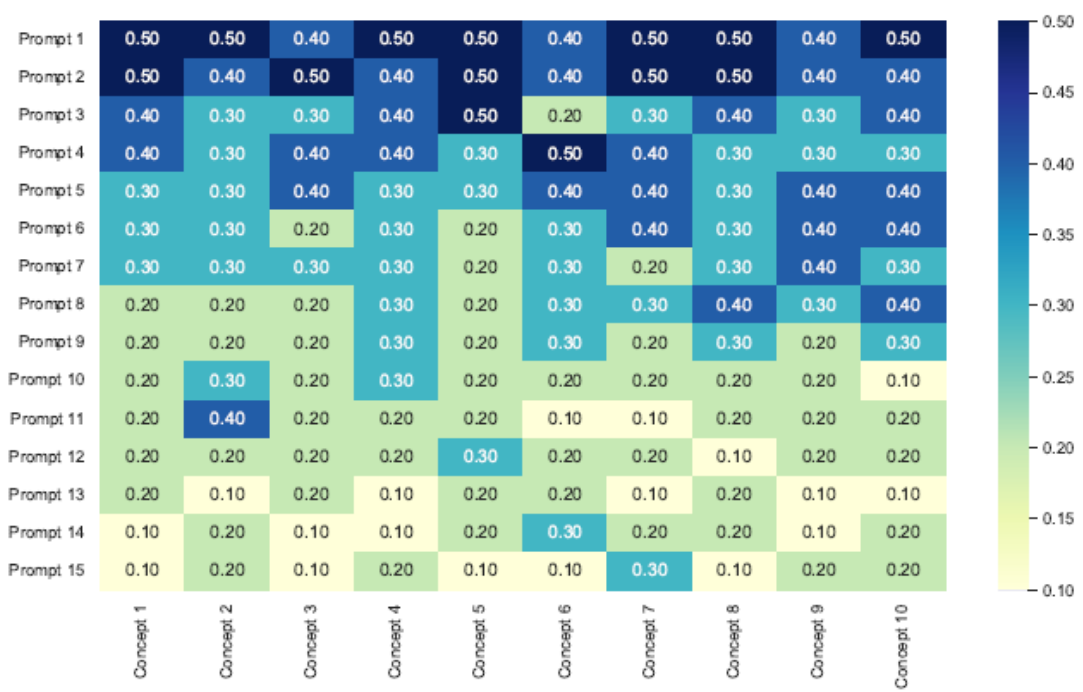
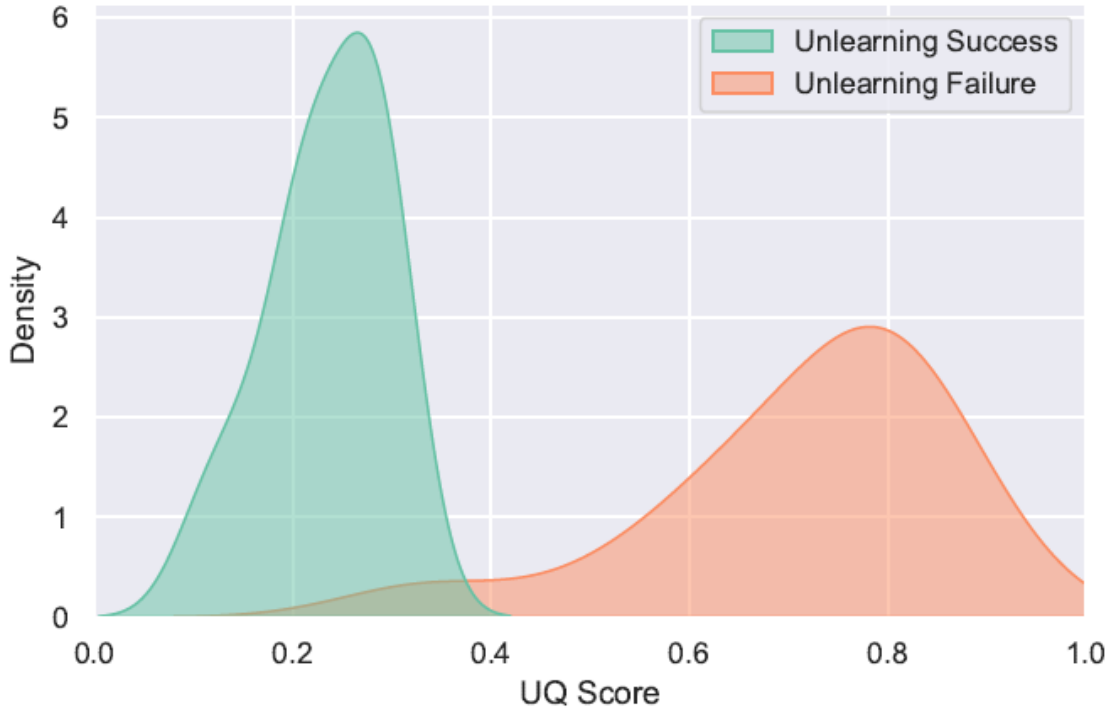


Illustration for the representation-level post-unlearning uncertainty quantification

Results





Heatmap of USR and UQ Scores (α = 0.5) for concept unlearning success and failure groups

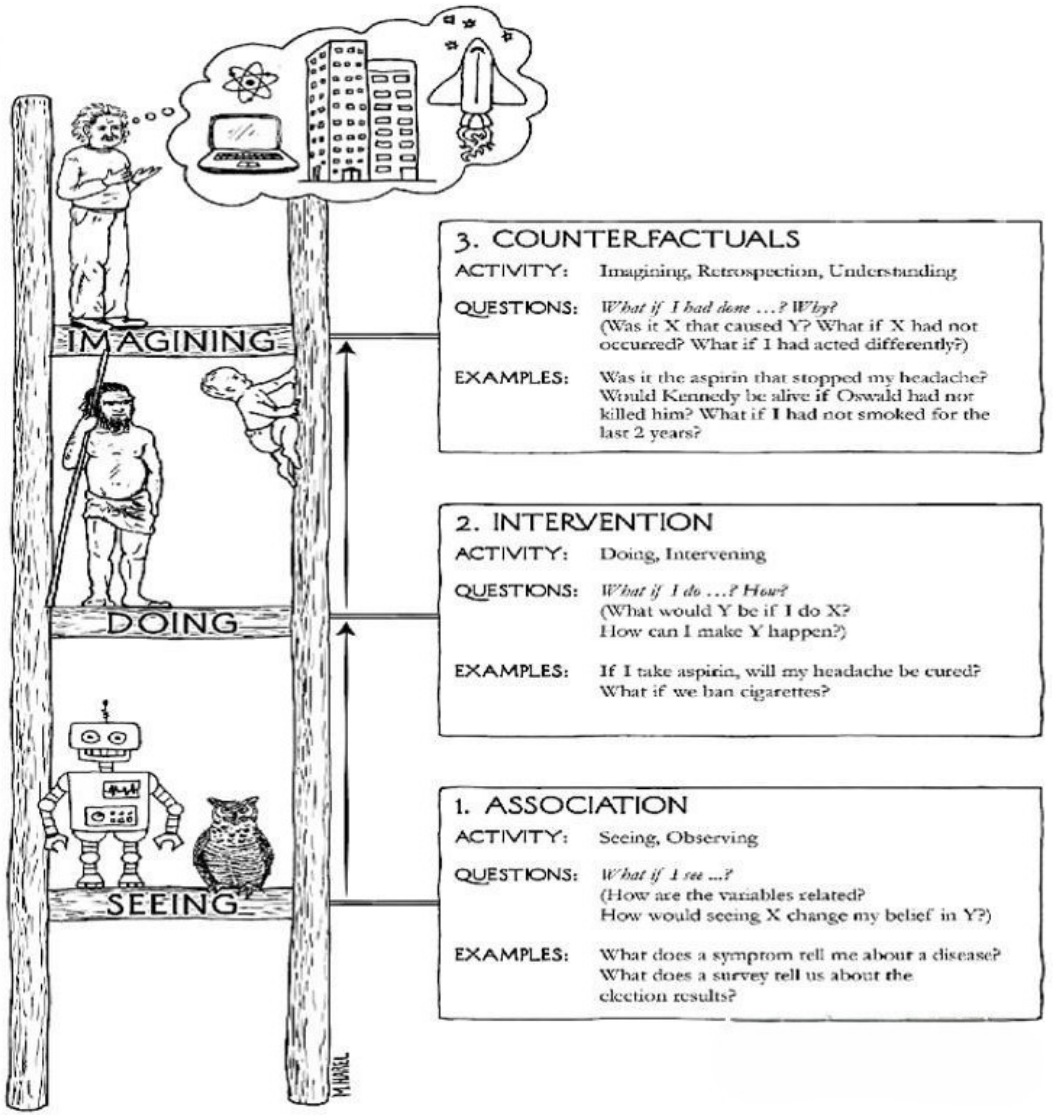
The Future

Future Work Plans

"Actual" Causality

"Causality-in-mean"

Statistics



The Ladder of Causation

Counterfactual learners, on the top rung, can imagine worlds that do not exist and infer reasons for observed phenomena.

Tool users, such as early humans, are on the second rung if they act by planning and not merely by imitation.

Most animals, as well as present-day learning machines, are on the first rung, learning from association.

Future Work Plans

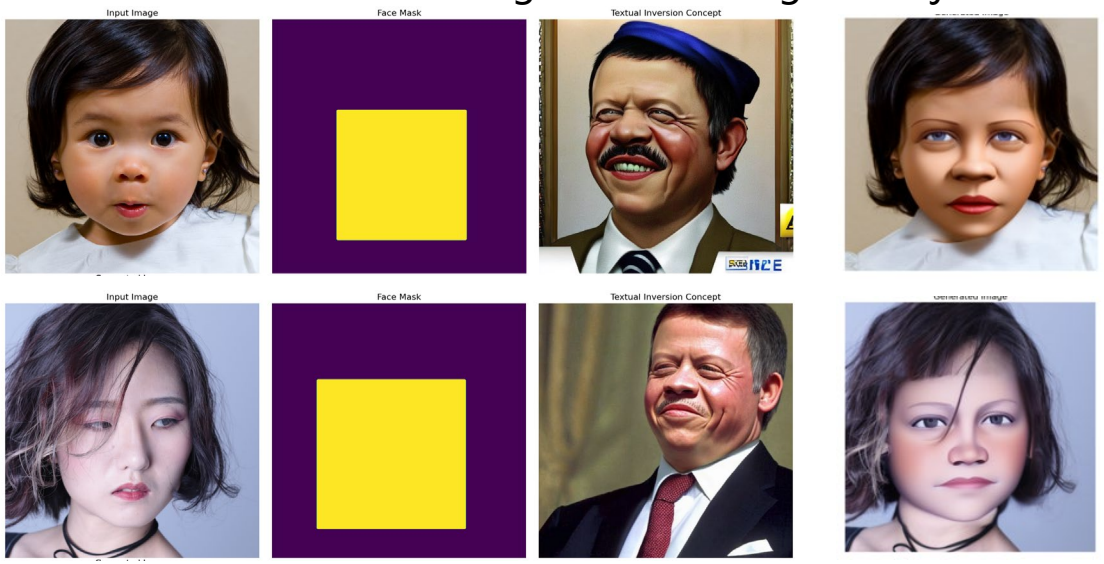
Causal Concept Diffusion

Source image | Concept generative image | Concept identity injection

The background area is also changed

Counterfactual Generation

Source image | Face mask | Concept generative image | Concept identity injection



- Preserving context
- Limiting the impact of identity embedding

By treating identity embedding as a causal intervening variable and intervening only in the latent space of the face, we can achieve **local**, **controllable**, and **explainable** counterfactual generation in the generated image.

The End

Thank You