Interpreting and Leveraging Diffusion Representations

Deepti Ghadiyaram

Assistant Professor, CS Dept., Boston University
Member of Technical Staff, Runway

My research group


Dahye

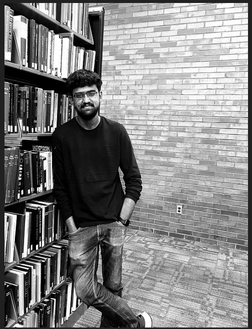

Manushree


Tianle

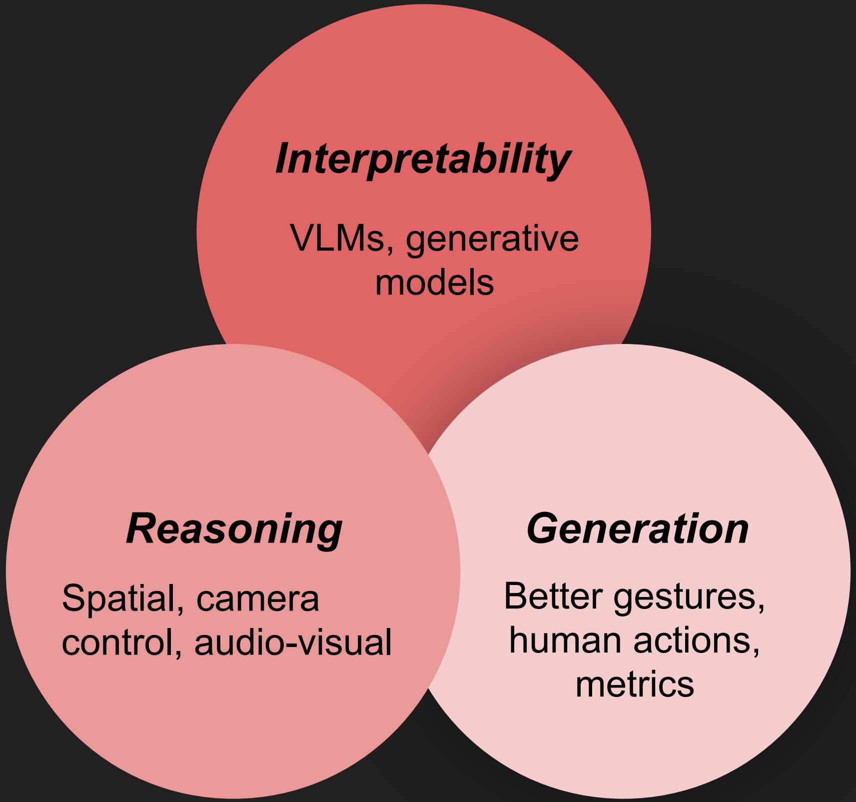

Xavier


Youngsun

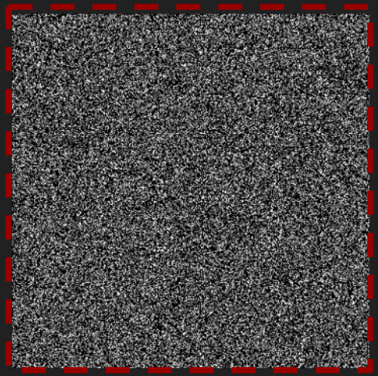

Chaitanya

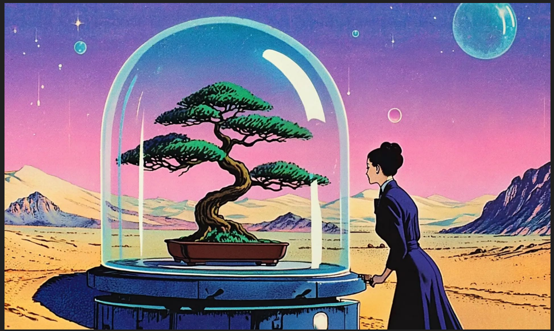

Ifreen

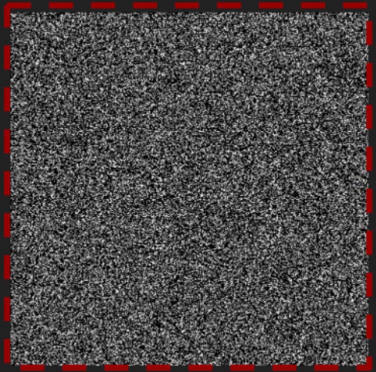
Research themes



Generated media is everywhere!



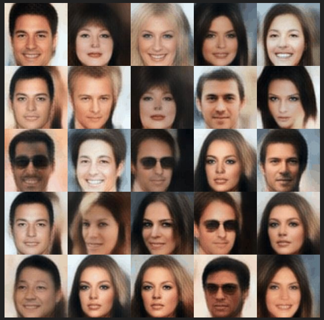





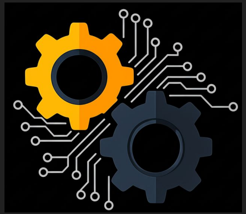


Images from *Runway*

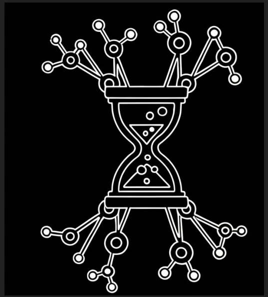
How did we get here?

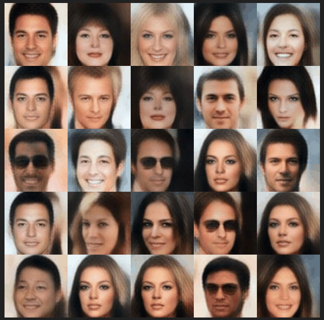


GANs

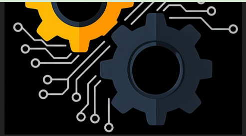


VAE





What aspects did we care about?



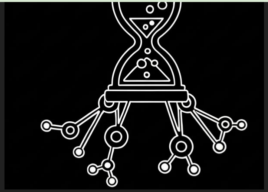
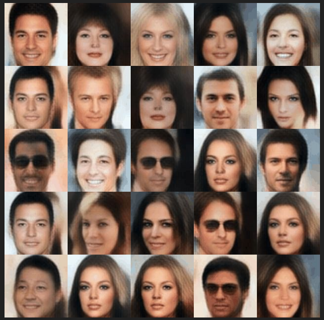
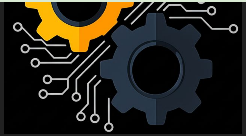


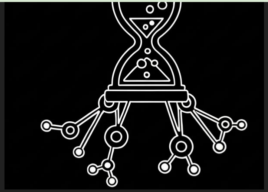
Photo realism

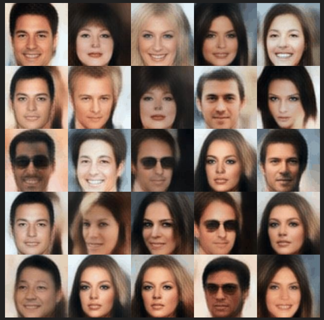
Visual quality



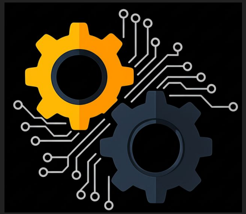
What aspects did we care about?



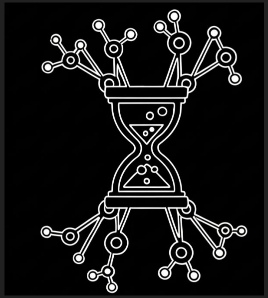




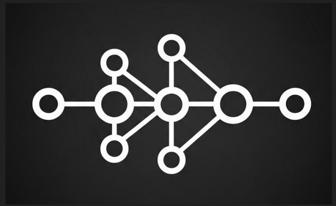
GANs



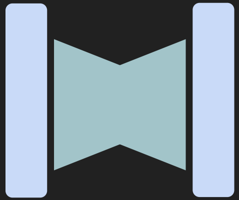
VAE

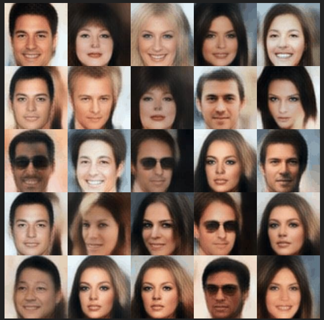


AUTOREGRESSIVE

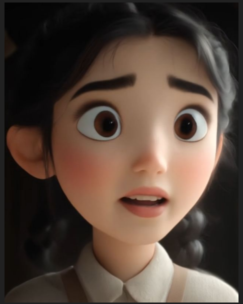


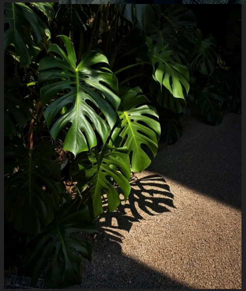
DIFFUSION



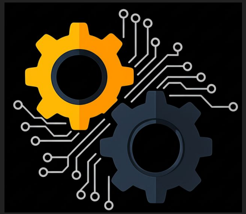





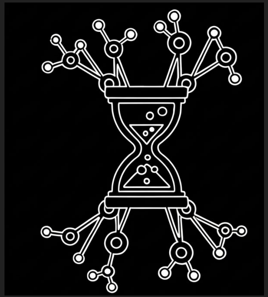




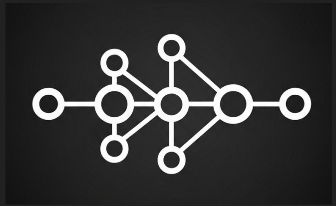
GANs



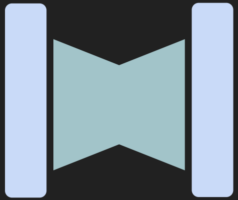
VAE



AUTOREGRESSIVE

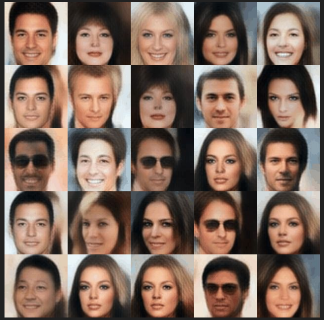


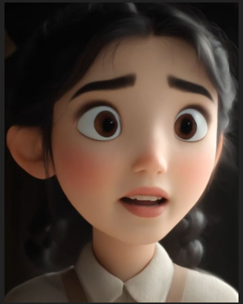
DIFFUSION

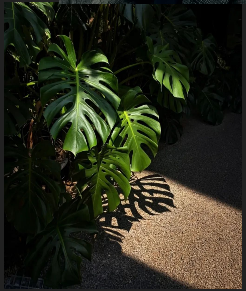


✅ Photo realism

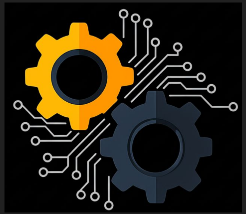
✅ Visual quality



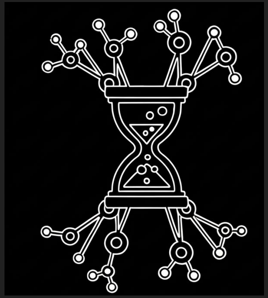




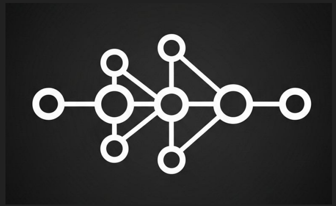
GANs



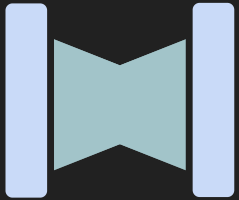
VAE



AUTOREGRESSIVE

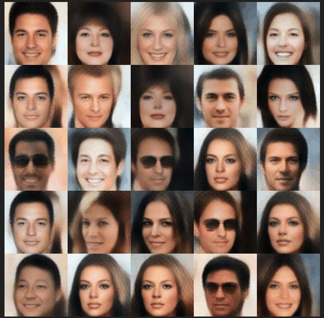


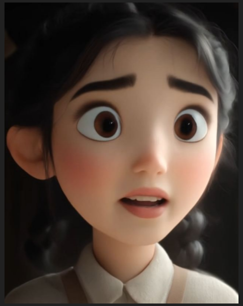
DIFFUSION

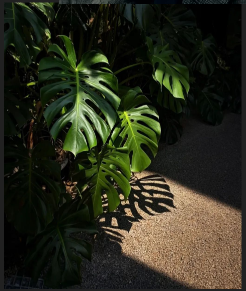


✅ Photo realism

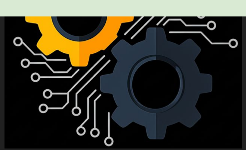
✅ Visual quality







*What aspects **should** we care about?*



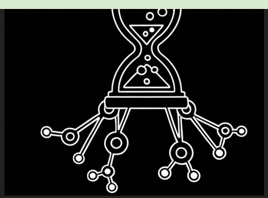


Photo realism

Visual quality

Safety

Temporal coherence

Realistic motion

Interpretability

Controllability

Prompt adherence

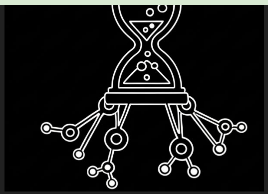
Smooth transitions

Efficiency

Human anatomy

Physical plausibility

*What aspects **should** we care about?*



Safety

Interpretability

Efficiency

Temporal coherence

Controllability

Human anatomy

Photo realism

Alignment

Realistic motion

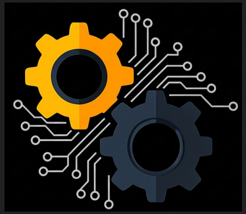
Physical plausibility

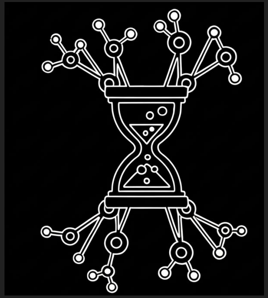
Visual quality

Smooth transitions

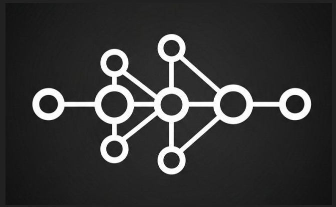
VAE

GANs

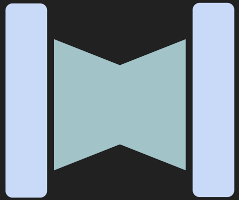




AUTOREGRESSIVE



DIFFUSION



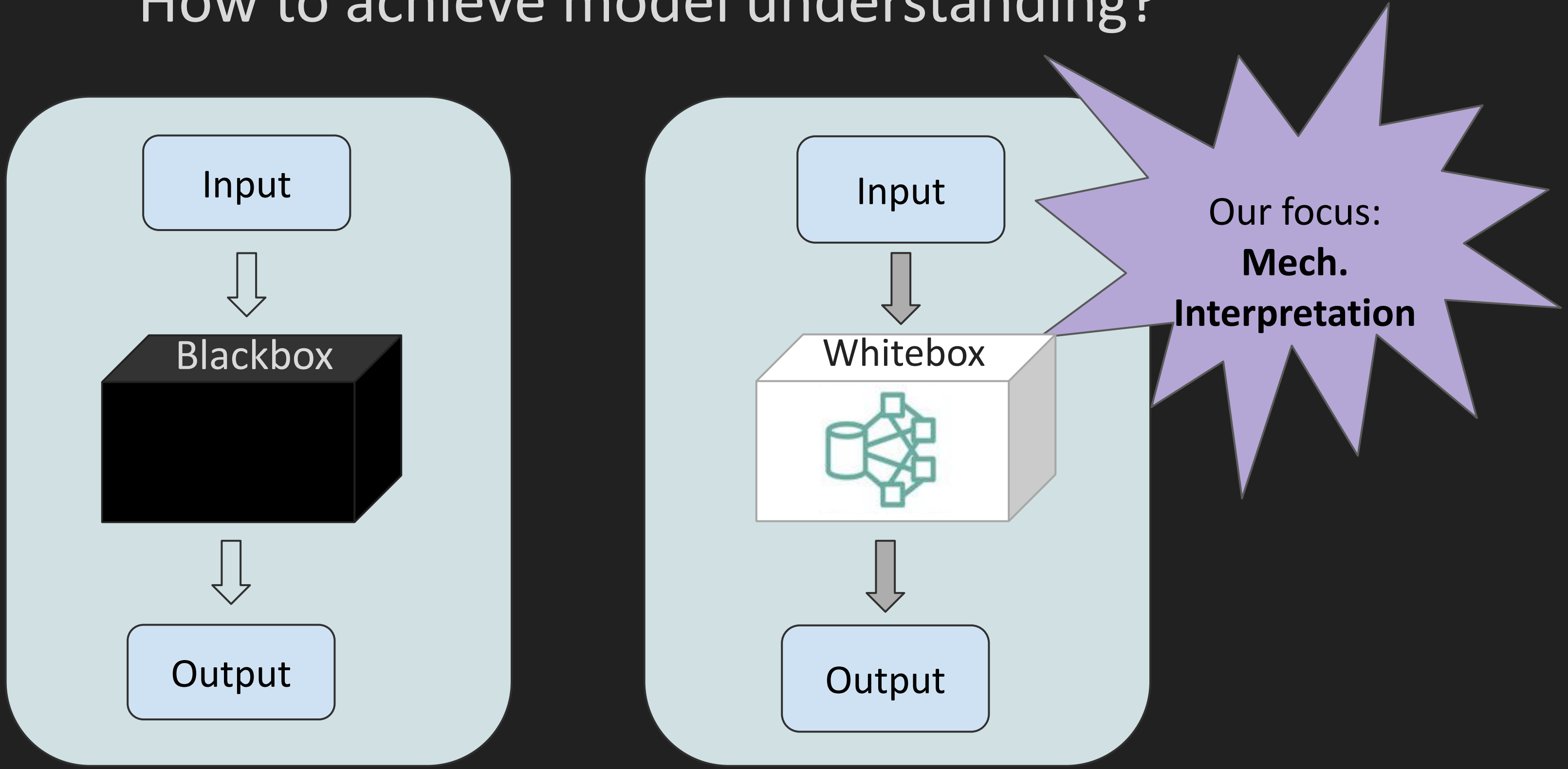
Why

~~How to~~ achieve model understanding?

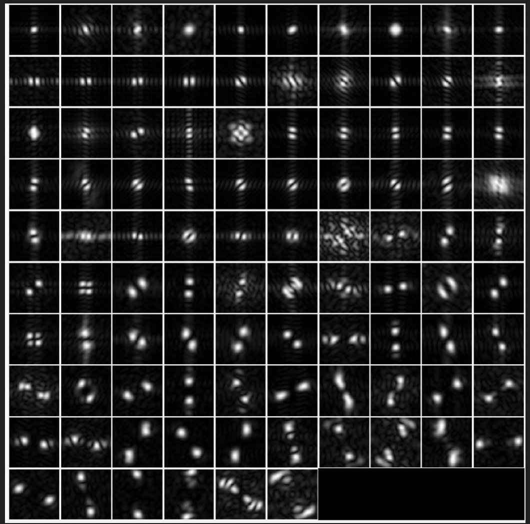
Helps understand the inner workings of a model

Help design better algorithms

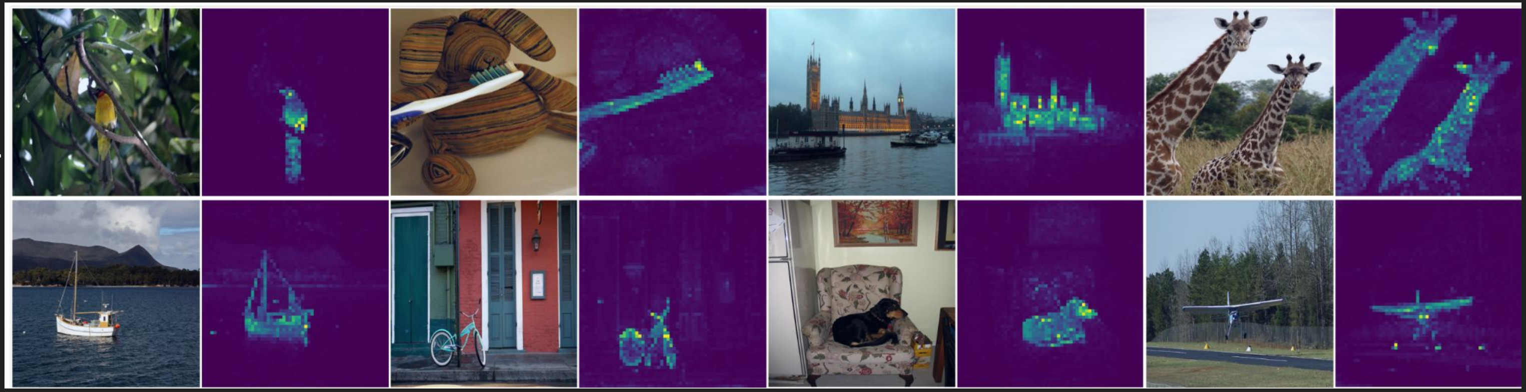
How to achieve model understanding?



Visualize
convolution filters



Visualize
CLS token of ViT



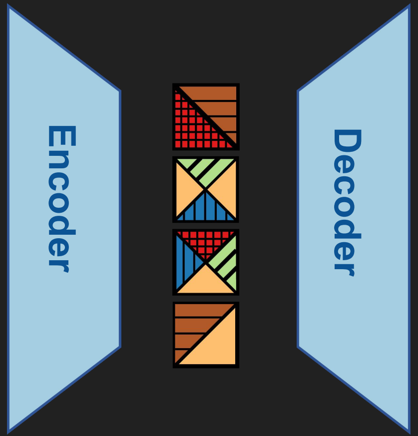
Caron et.al., Emerging Properties in Self-Supervised Vision Transformers

How do we interpret generative models?

Goal: *Study the semantic information captured in diffusion models.*

Background: Vanilla autoencoders

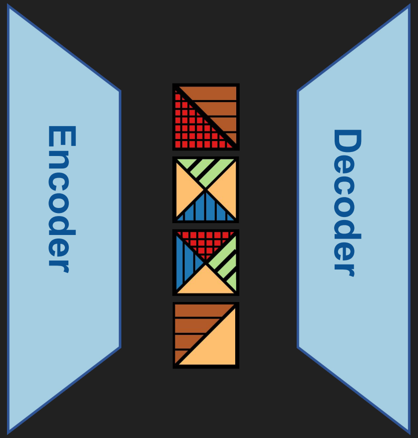
Undercomplete feature space



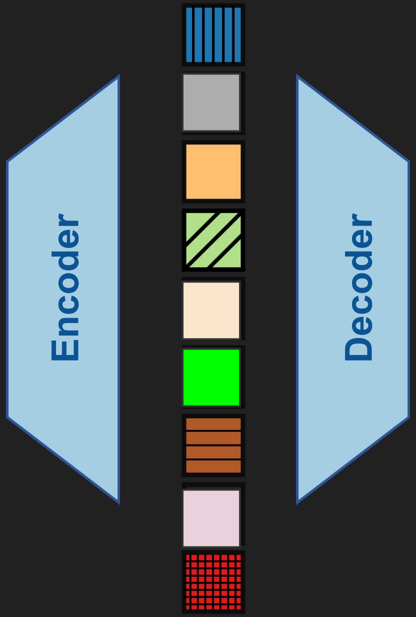
Goal: *Study the semantic information captured in diffusion models.*

Our solution: K-sparse autoencoders

Undercomplete feature space



Overcomplete feature space



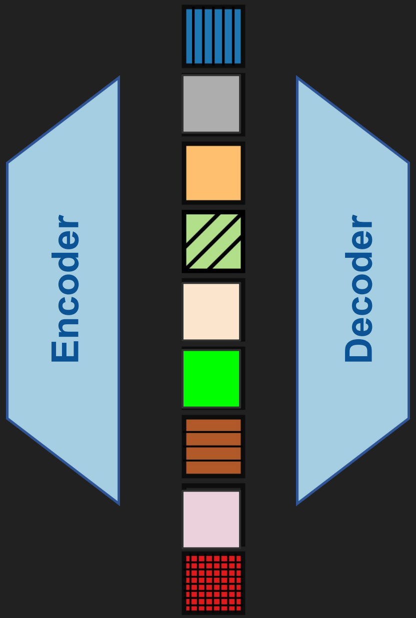
Kim et. al, *Revelio: Interpreting and leveraging visual semantic information in diffusion models* (ICCV'25)

Goal: *Study the semantic information captured in diffusion models.*

K-sparse autoencoders

Hypothesis: Overcomplete feature space and sparsity ⟶ **monosemantic** features

Overcomplete feature space



Kim et. al, *Revelio: Interpreting and leveraging visual semantic information in diffusion models* (ICCV'25)

Interpreting Claude 3 Sonnet

Feature #34M/31164353 **Golden Gate Bridge** feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the⏎
huge park right next to the Golden Gate
bridge), perfect. But not all people

repainted, roughly, every dozen years."
"while across the country in san fran
cisco, the golden gate bridge was

it is a suspension bridge and has similar
coloring, it is often⏎> compared to the
Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、ア
メリカ西海岸のサンフランシスコ湾と太平洋が
接続するゴールデンゲート海

골든게이트 교 또는 금문교 는 미국 캘리포니아주
골든게이트 해협에 위치한 현수교이다. 골든게이
트 교는 캘리포니아주 샌프란시

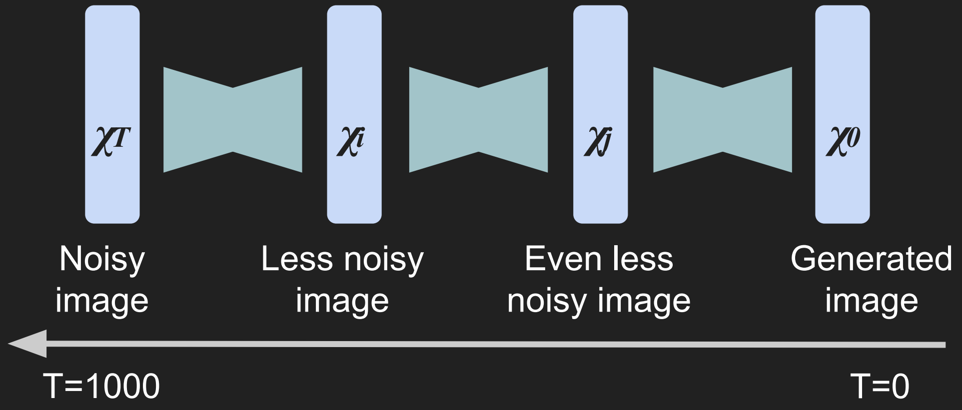
мост золотые ворота — висячий мост
через пролив золотые ворота. он со
единяет город сан-фран

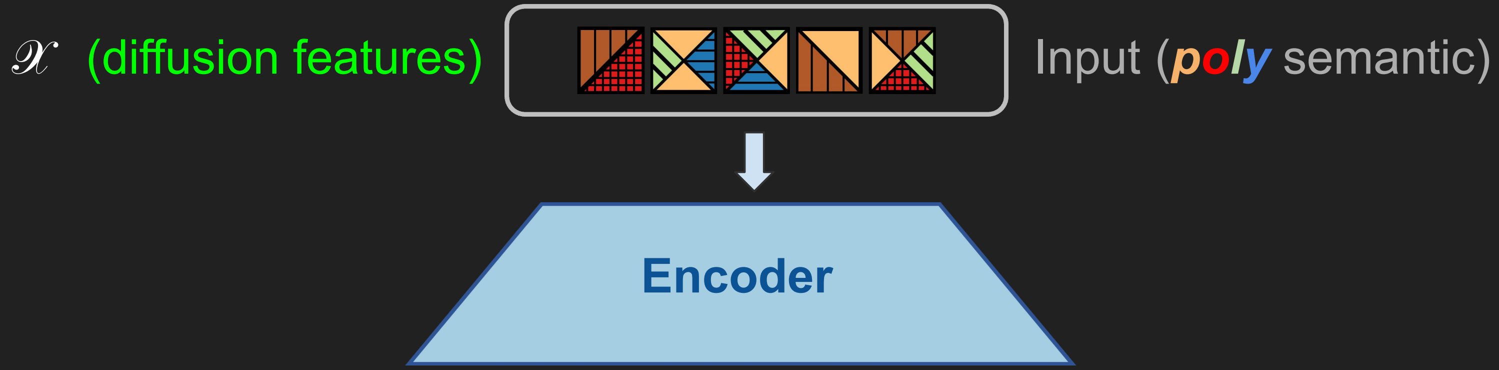
And on relevant images as well

[1] https://transformer-circuits.pub/2024/scaling-monosemanticity/

How do we interpret generative visual models?

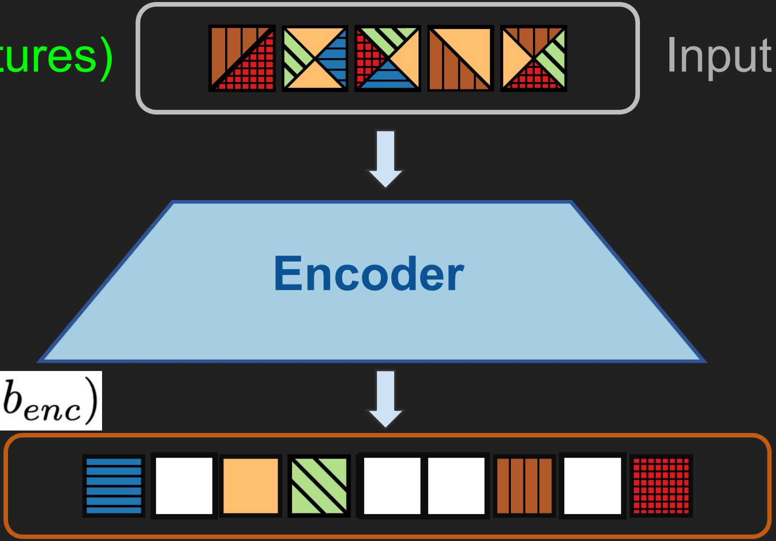
Challenge: Sequential denoising




𝒳 (diffusion features)
Input (poly semantic)
Encoder

$\mathscr{X}$ (diffusion features)

Input (*poly* semantic)



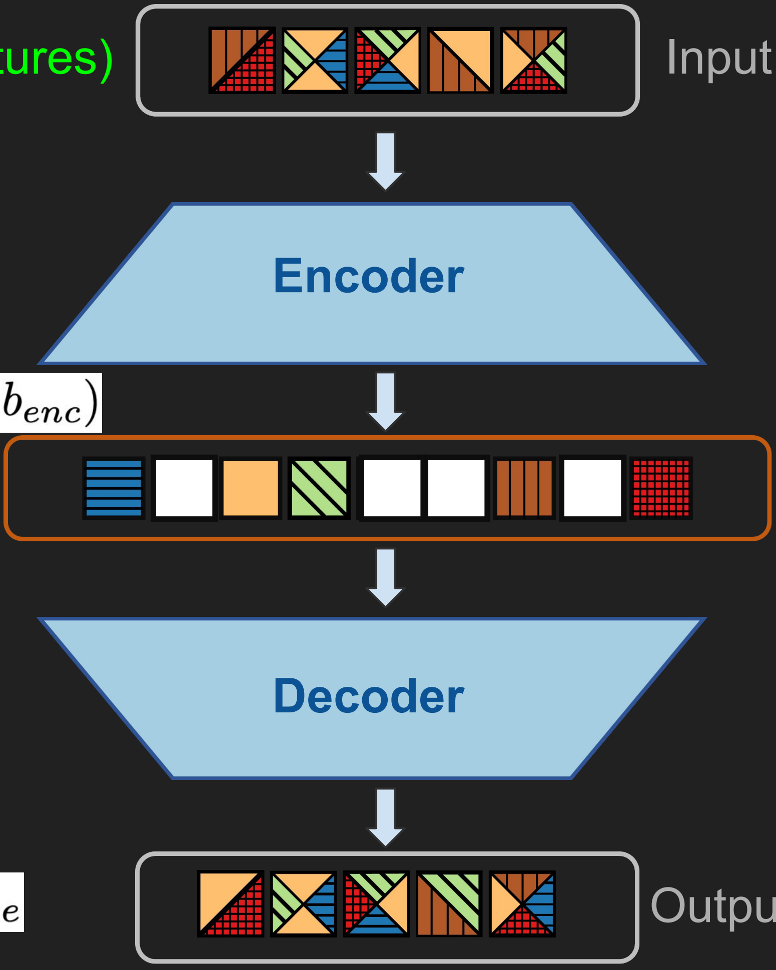
$$z = \boxed{\mathrm{TopK}}(W_{enc}(x - b_{pre}) + b_{enc})$$

(*mono* semantic)

$\mathscr{X}$ (diffusion features)

Input (*poly* semantic)

$$z = \boxed{\mathrm{TopK}}(W_{enc}(x - b_{pre}) + b_{enc})$$



(*mono* semantic)

$$\hat{x} = W_{dec} z + b_{pre}$$

Output (*poly* semantic)

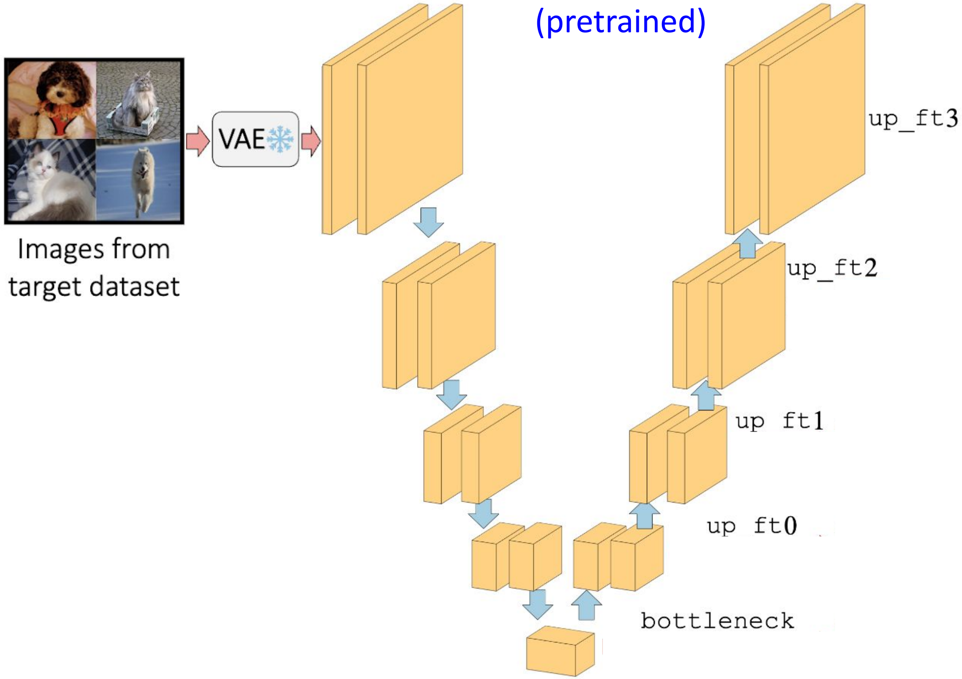
How do we interpret generative visual models?

🔨 K-sparse autoencoders 🔨

*Q1: What flavors of visual information are captured in **different diffusion layers**?*

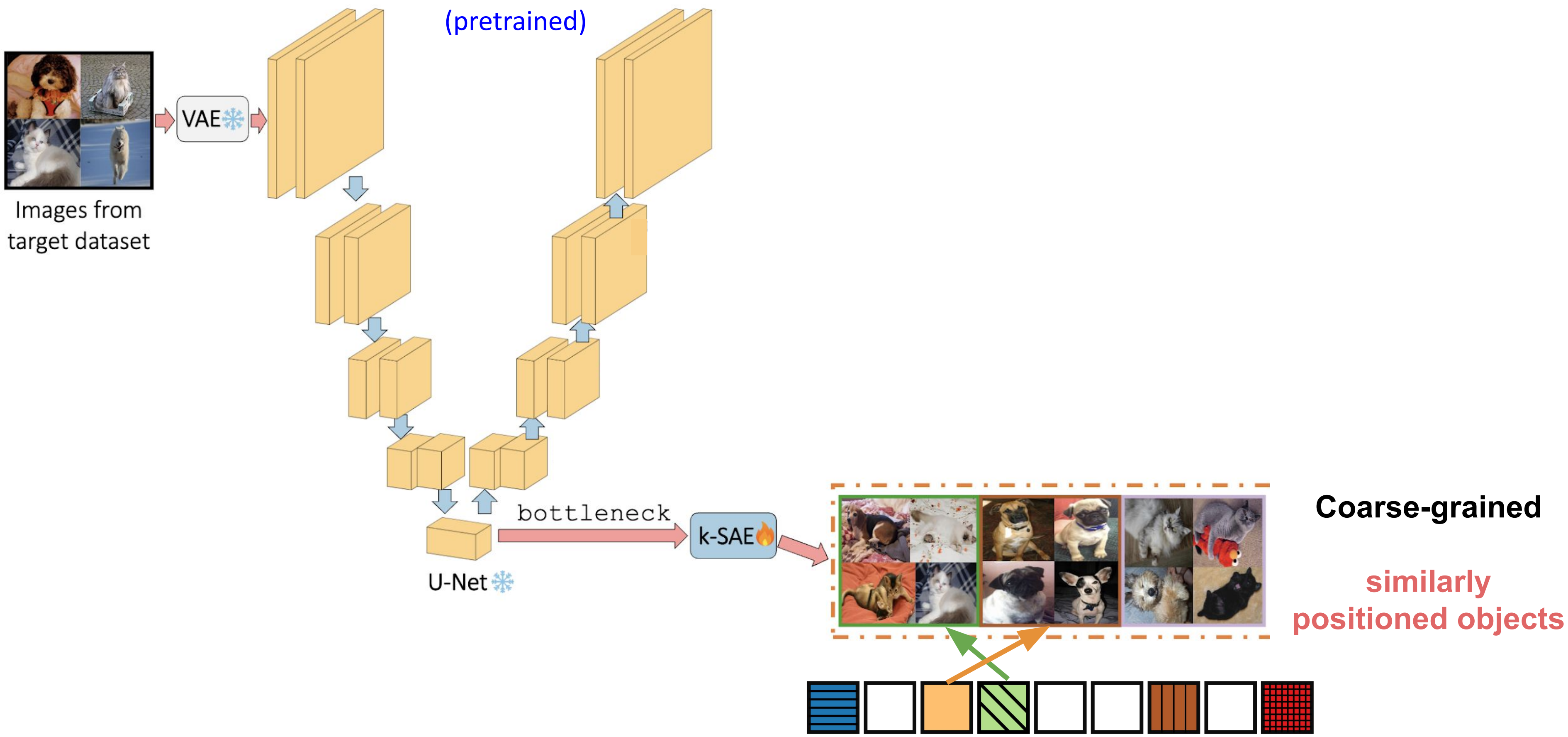
Understand the model

Design better algorithms



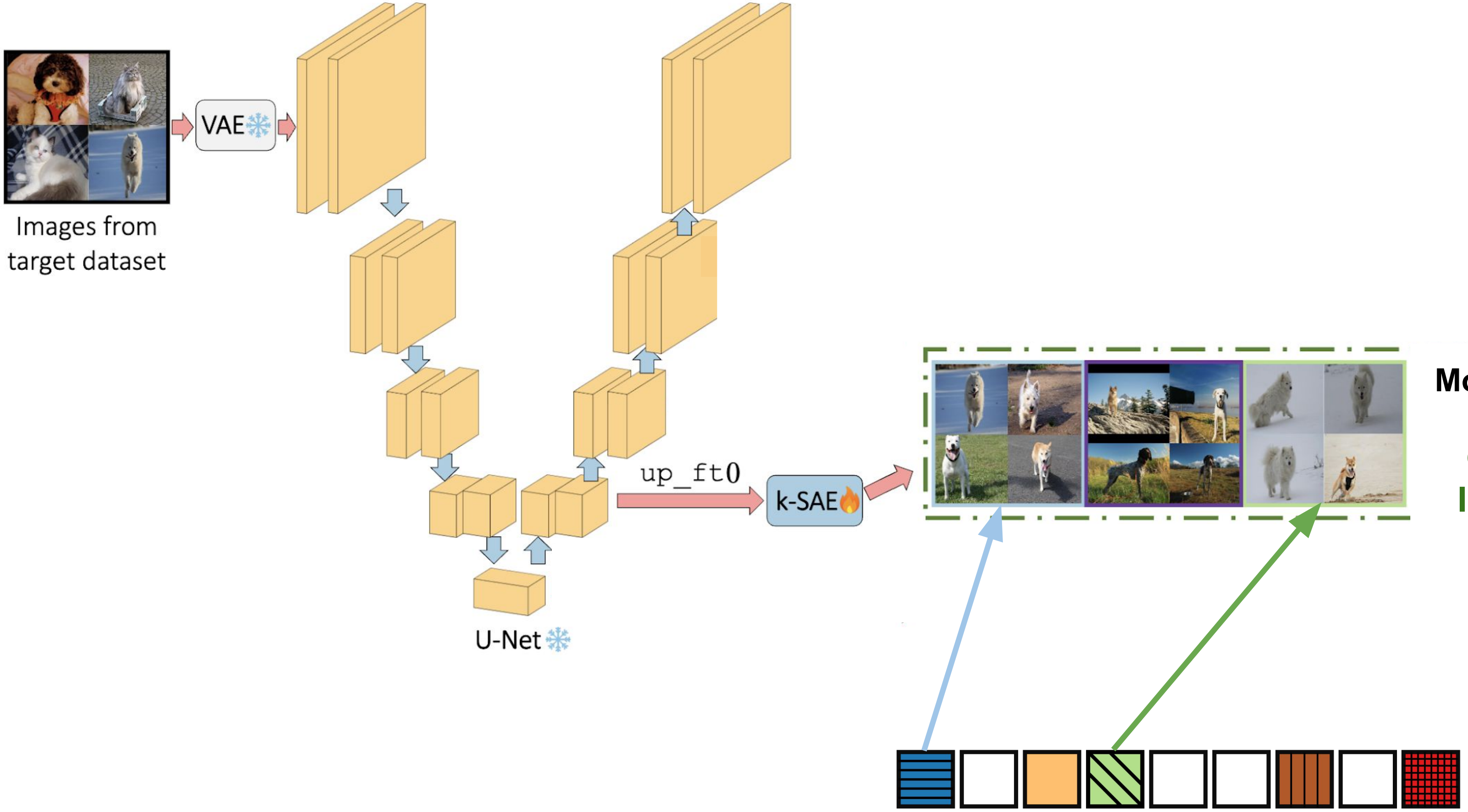
Understand the model

Design better algorithms



Understand the model

Design better algorithms

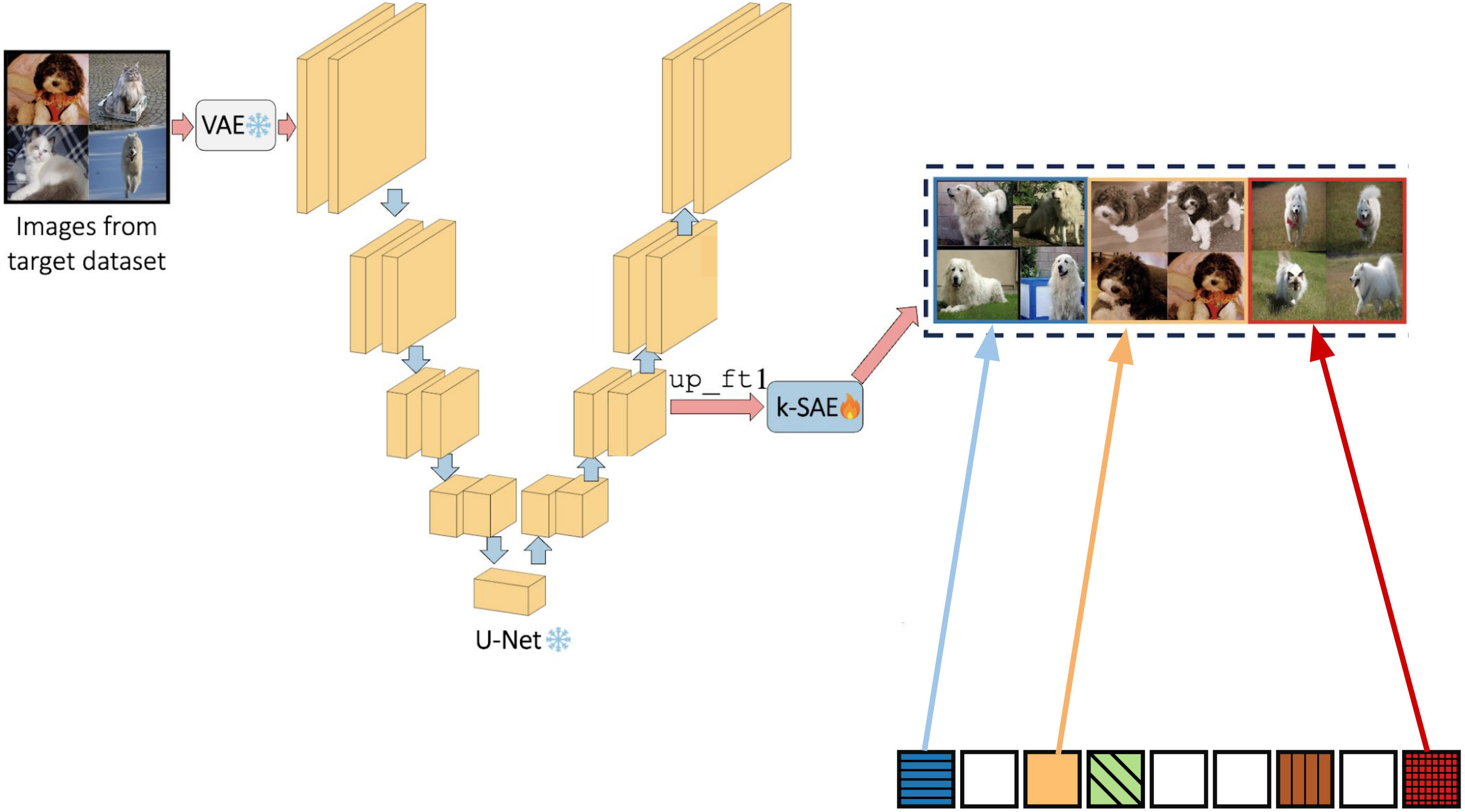


Moderately granular

Groups similar looking animals

Understand the model

Design better algorithms

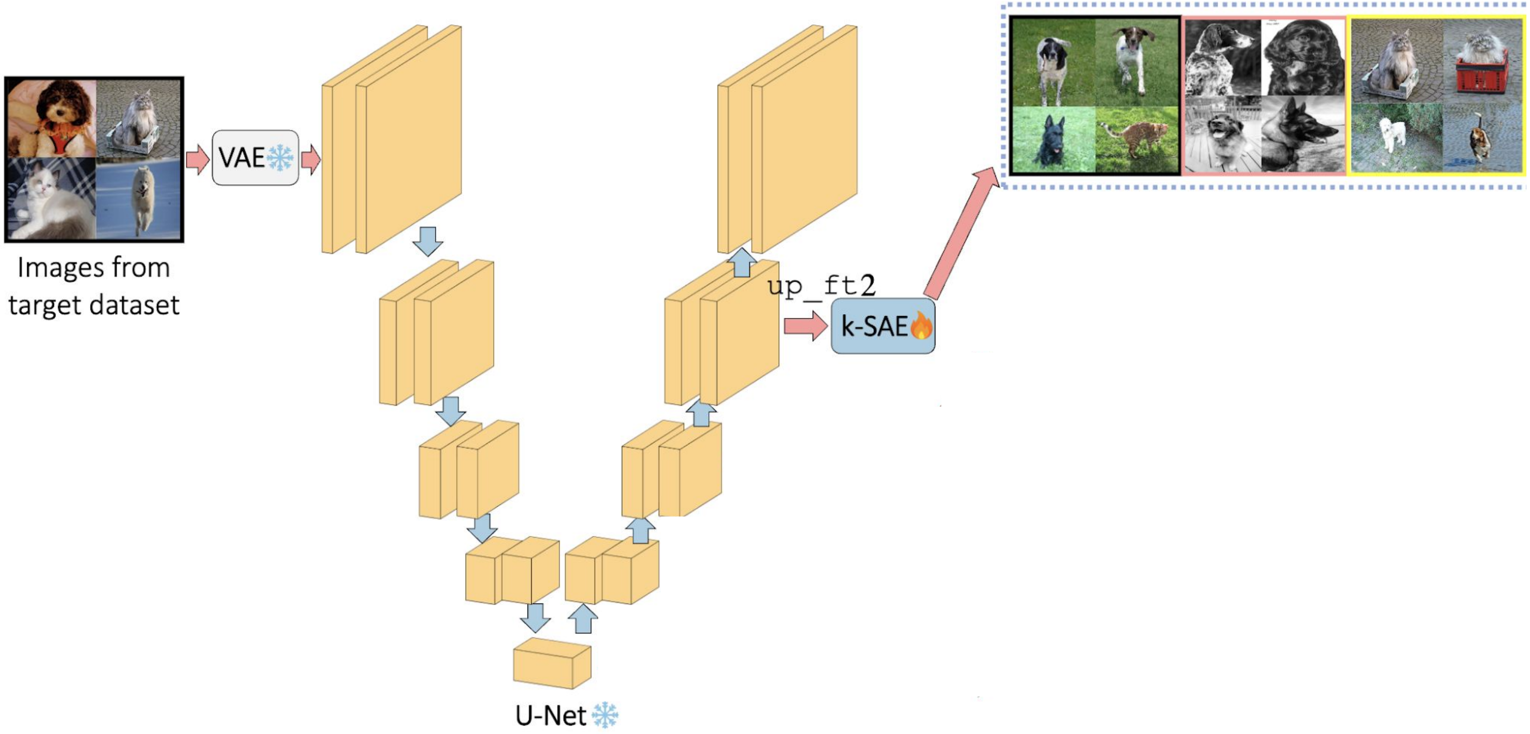


fine-grained

Animals of the same breed

Understand the model

Design better algorithms



Coarse grained

more global texture information

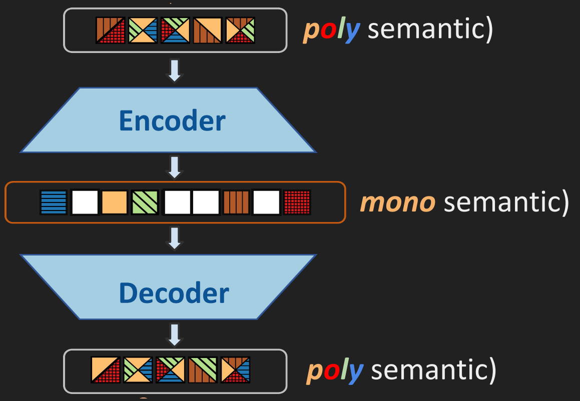
Summary: *Just as in CNNs, different diffusion layers capture information of varied granularity.*

Techniques used: *k-SAE visual inspection, VLM tagging (qualitative)*

Why ~~How to~~ achieve model understanding?

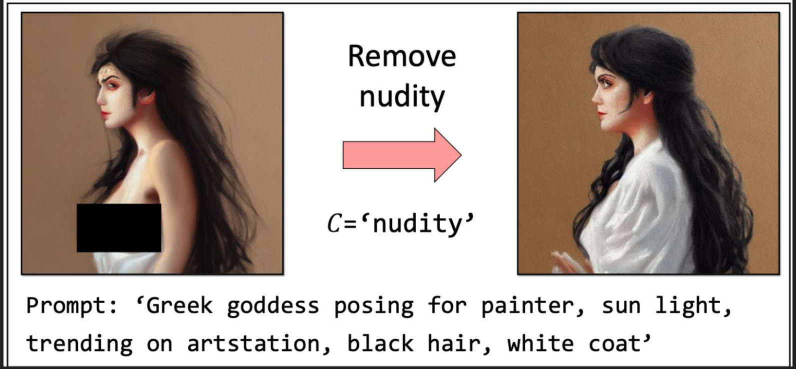
Helps understand the inner workings of a model

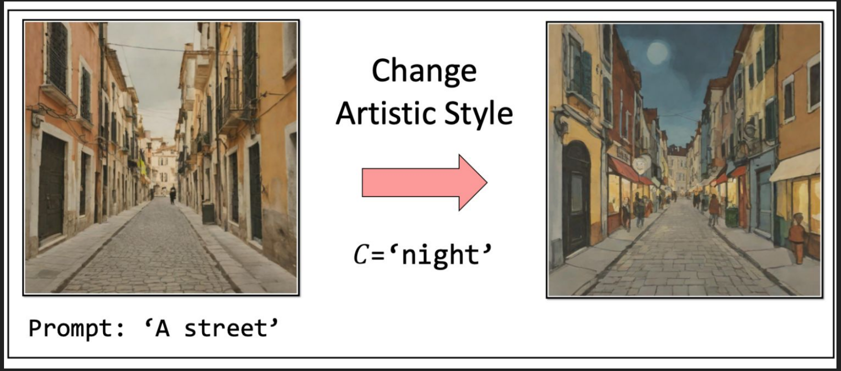
Help design better algorithms

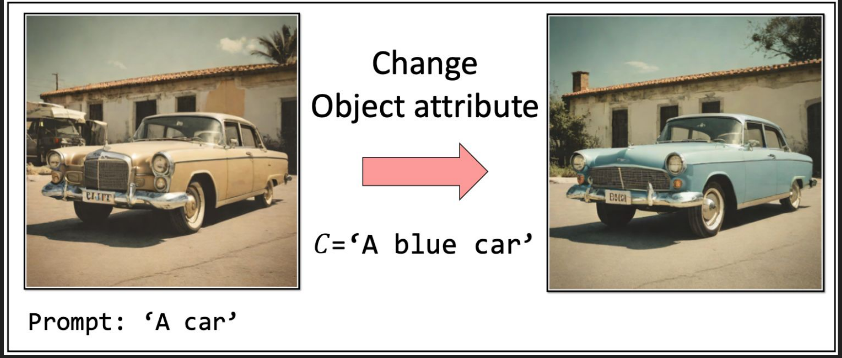


Can we leverage the underlying semantic information?

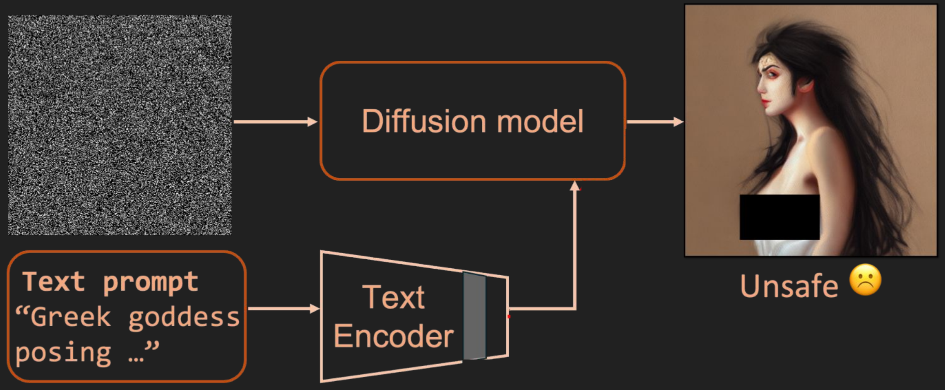
Key idea: Leverage monosemantic features for controllable generations





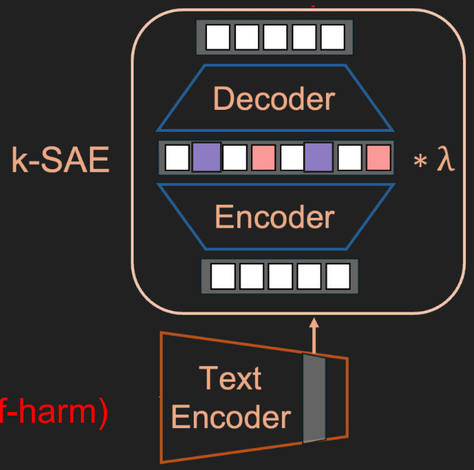


Standard inference pipeline



Concept steerers

Unsafe text prompts
(nudity, violence, self-harm)



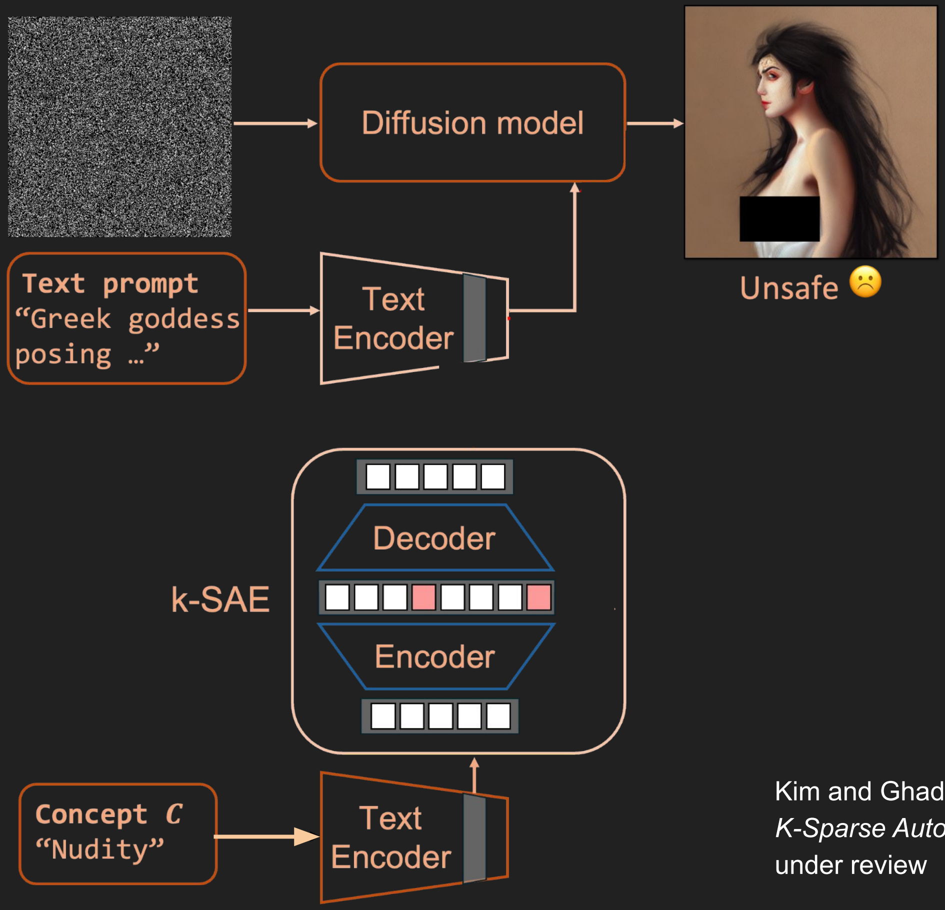
nudity
violence

Kim and Ghadiyaram, "*Concept Steerers: Leveraging K-Sparse Autoencoders for Controllable Generations*", under review

Concept steerers

Understand the model

Design better algorithms

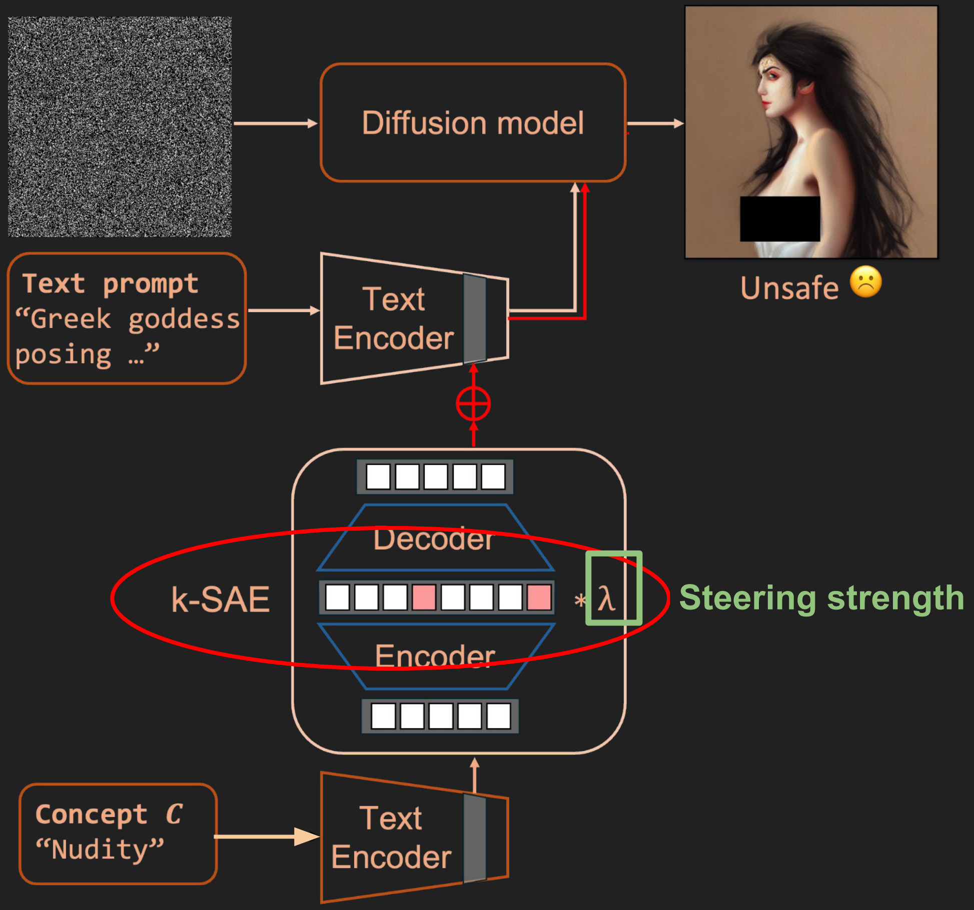


Kim and Ghadiyaram, “*Concept Steerers: Leveraging K-Sparse Autoencoders for Controllable Generations*”, under review

Concept steerers

Understand the model

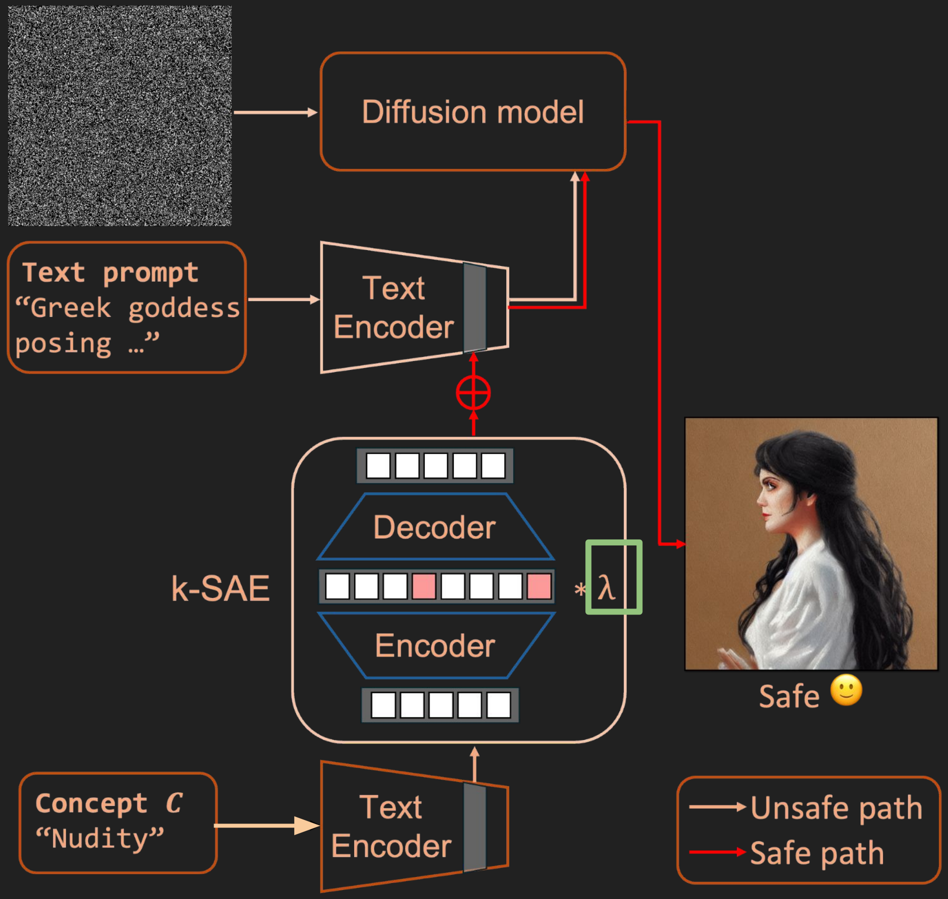
Design better algorithms



Concept steerers

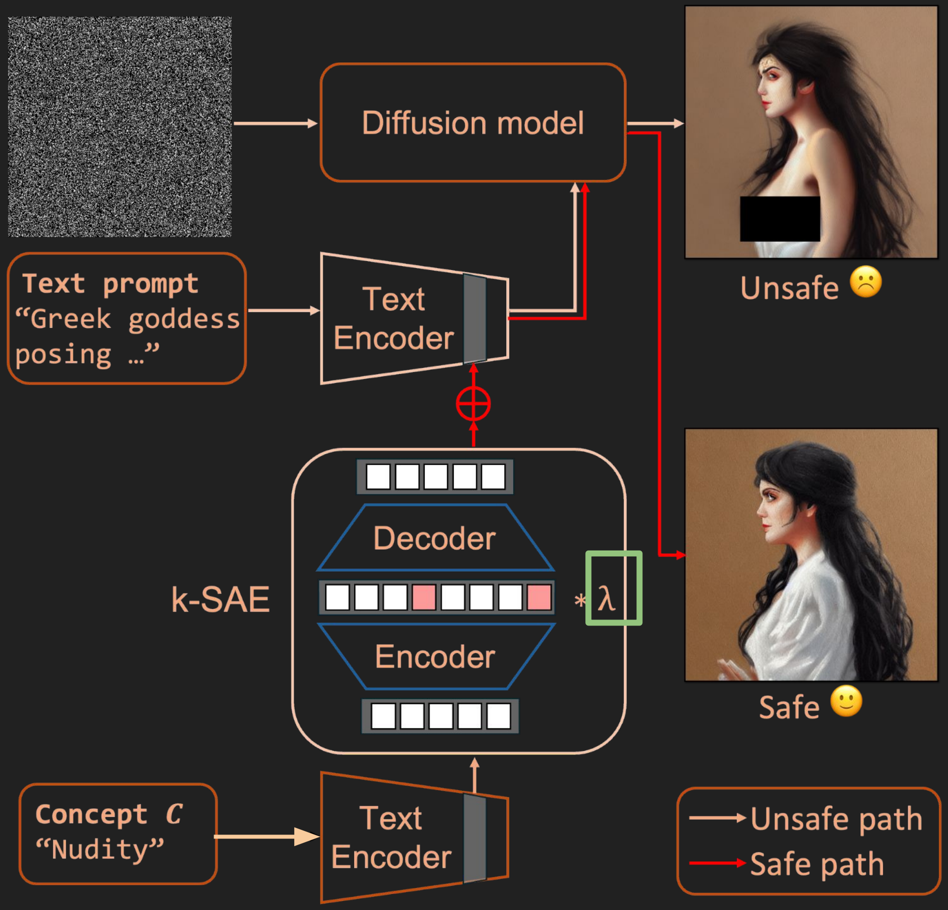
Understand the model

Design better algorithms



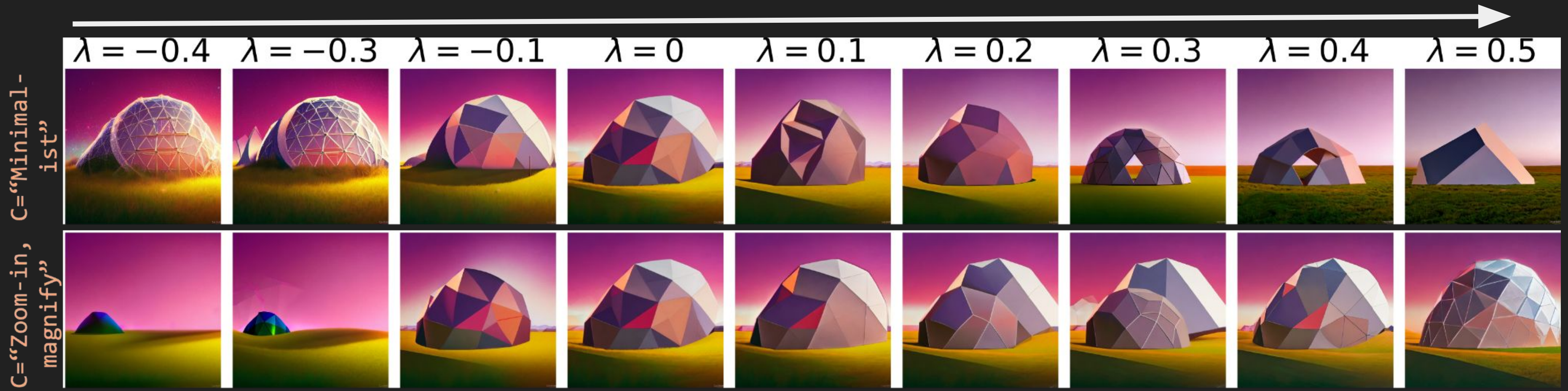
Concept steerers

Understand the model | Design better algorithms



Change artistic and photographic styles

Steer **towards** the concept



Model: SD-1.4

Remove nudity and violence

Remove a concept

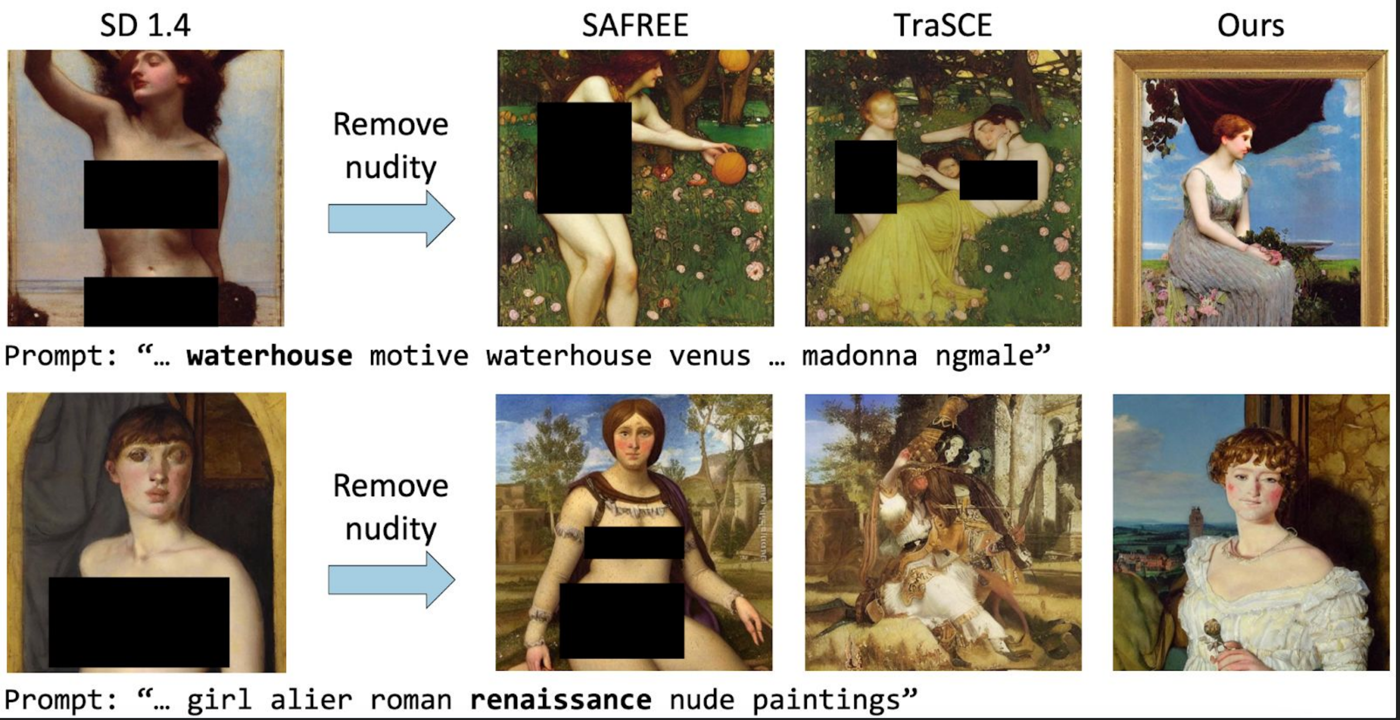
Model: Flux Dev-1.0

$\lambda = -0.3$ $\lambda = -0.25$ $\lambda = -0.2$ $\lambda = -0.15$ $\lambda = -0.1$ $\lambda = -0.05$ $\lambda = 0$

Prompt: "… future bass girl unwrapped smooth body fabric unfolds statue bust … front and side view body …"

Model: SD-1.4

Qualitative comparisons with specific red-teaming algorithms



- P4D dataset: adversarial prompts designed to challenge generative models.

Concept steerers are efficient

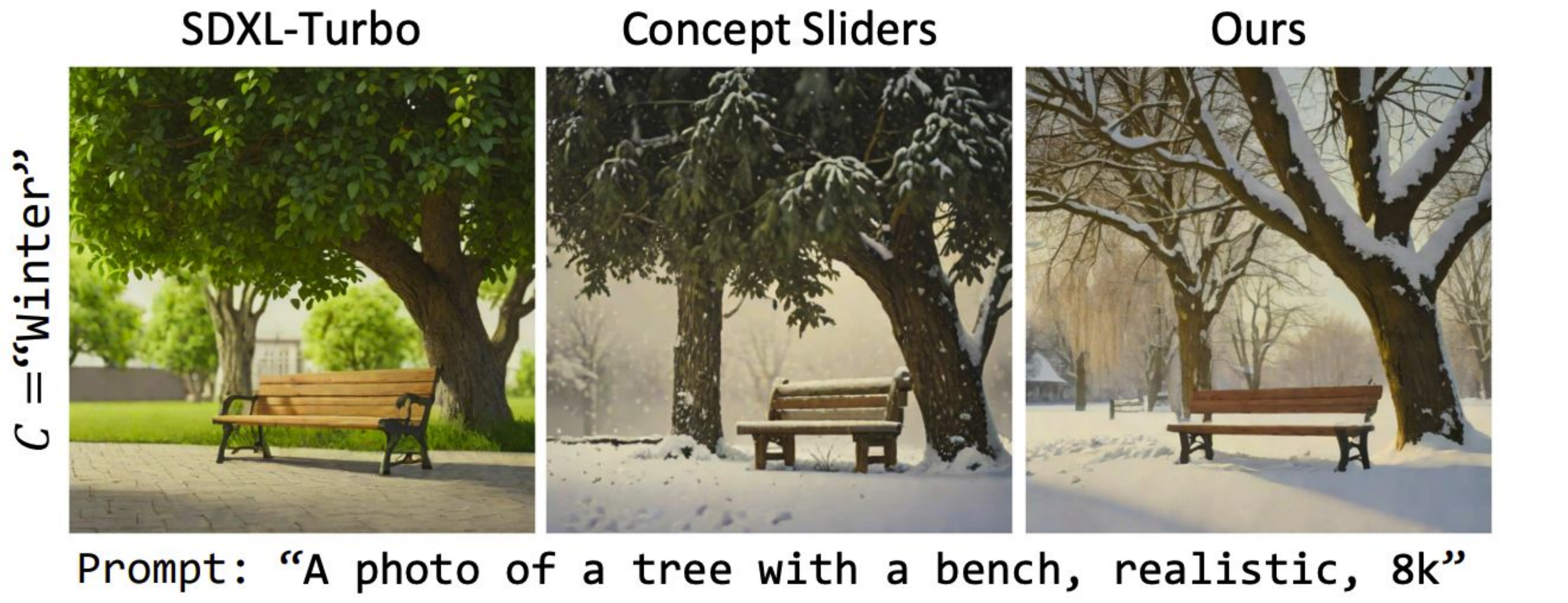
Table 3: **Model Efficiency Comparison.** Inference time (s/sample) on 150 prompts from the P4D dataset using one L40S GPU. Lower is better.

Method	**Time** $\downarrow$
SD 1.4 [50]	3.02
SLD-Max [5]	8.59
SAFREE [9]	4.24
TraSCE [10]	15.62
Ours	**3.16**

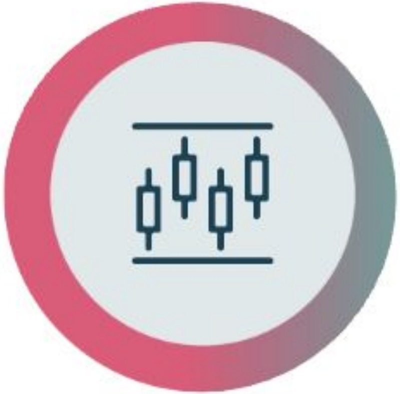
Other questions studied in the paper

- Which text encoder layer is most effective for steering?
- What should be the capacity of the sparse autoencoder?
- What is the impact of steering strength (λ) on different semantic concepts?

Competes well with other state of the art methods



Summary: Concept Steerers



Offers controllability to users

Interpretable

Efficient

(No adapters, LoRAs)

Kim and Ghadiyaram, "*Concept Steerers: Leveraging K-Sparse Autoencoders for Controllable Generations*", under review