From Interpretability to Control
Steering Diffusion Models with Concept Directions

Hang Li

PhD, LMU Munich

Multimodal generation and interpretability

Research Scientist, Meta London

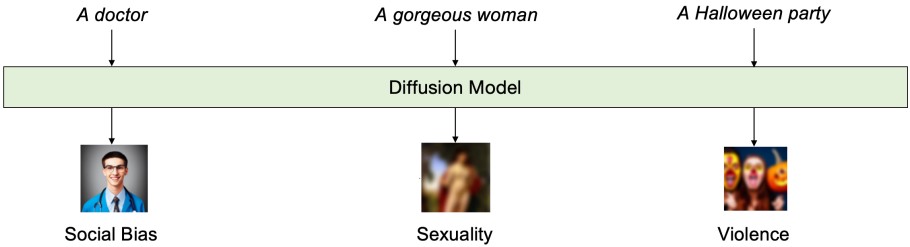
Video foundation models and controllable generation

Outline

Text-to-Image Diffusion Models Can Fail Silently

Even neutral prompts can trigger biased or harmful outputs. Where do these failures originate?

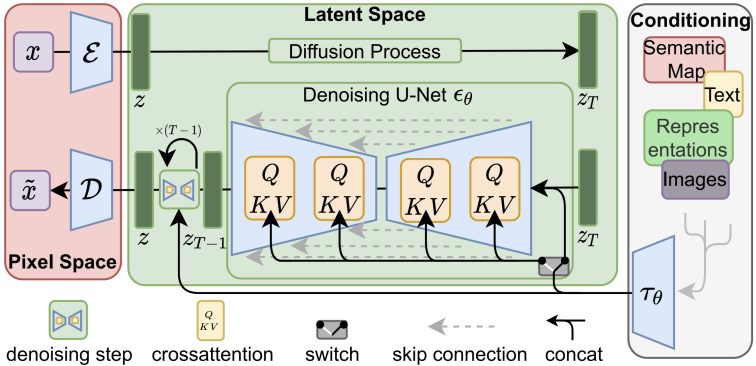
Can we locate the activations that represent harmful attributes and disable them at inference time to enable responsible generation?



Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Which Units Cause a Given Output Attribute?

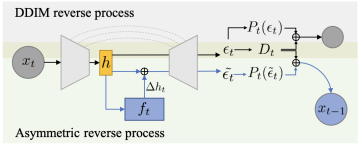
The U-Net bottleneck (h-space) concentrates semantic information into a structured, low-dimensional manifold.

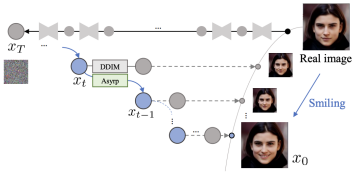


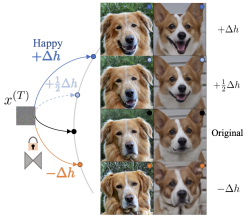
Rombach et al., CVPR 2022

h-space: The Semantic Bottleneck

Meaning is encoded as *directions* in the *h*-space. Adding a vector Δh to the bottleneck therefore *shifts a semantic attribute*.







The intuition is the same as in word embeddings: $\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$.

Kwon et al., ICLR 2023

How to Find Directions for Any Concept?

No prior method discovers vectors for arbitrary concepts, especially the unsafe ones we most need to control.

Approach	External classifier	Human labels	Any concept
Unsupervised (PCA)	no	no[†]	no
Supervised	yes	yes	limited
Ours	**no**	**no**	**yes**

[†]Post-hoc human interpretation required; the target concept may not appear.

Observation: Text-to-image diffusion models are trained on aligned image–text pairs, giving them semantic associations for free: We can leverage this internalized knowledge to discover concepts

Method: Self-Supervised Concept Discovery

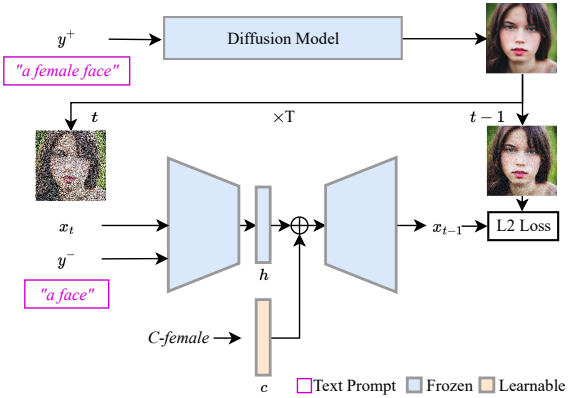
Key Idea: Learn a vector Δh in h-space that captures a target concept using only prompt pairs.

1. Generate with y^+ (*"a female face"*)
2. Remove concept: y^- (*"a face"*)
3. Optimize Δh to reconstruct

$$\Delta h^* = \arg\min_{\Delta h} \left\| \epsilon - \epsilon_\theta(x_t, t, \tau(y^-), h + \Delta h) \right\|^2$$

U-Net frozen; only Δh is learned.

Although learned from just one fixed prompt, the concept vector transfers to arbitrary prompts!

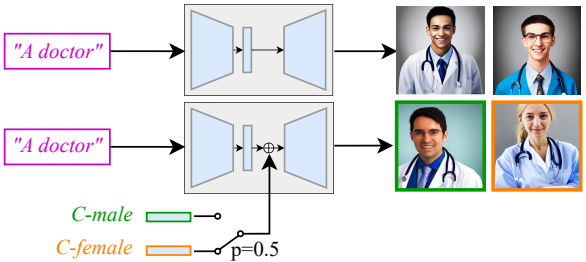


Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Fair Generation

Sample a learned attribute vector with equal probability at inference:

$\mathrm{h} \leftarrow \mathrm{h} + \Delta \mathrm{h}_g$, where $\Delta \mathrm{h}_g \in \{\Delta \mathrm{h}_\mathrm{m}, \Delta \mathrm{h}_\mathrm{f}\}$ with $p = 0.5$.

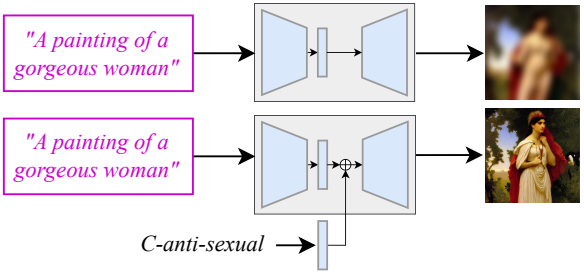


- One vector trained on *"person"*
- Transfers to all professions
- No retraining, no per-profession tuning

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Safe Generation

Add an anti-concept vector to suppress harmful content at inference:
$\mathbf{h} \leftarrow \mathbf{h} + \Delta\mathbf{h}_s$.



Adding $\Delta\mathbf{h}_{\textit{anti-sexual}}$ suppresses unsafe content

- y^+ contains unsafe concept, y^- removes it; recover $-\Delta\mathbf{h}$
- Stacks on top of existing safety methods (SLD, ESD)

SOTA on Winobias

A single self-supervised vector achieves state-of-the-art fairness across all professions.

Deviation ratio ($\downarrow$ = fairer)

	Gender			Gender+		
	SD	UCE	**Ours**	SD	UCE	**Ours**
Analyst	0.70	0.20	**0.02**	0.54	0.04	**0.02**
CEO	0.92	0.28	**0.06**	0.90	0.58	**0.06**
Teacher	0.30	0.06	**0.04**	0.48	0.16	**0.10**
Winobias	0.68	0.22	**0.17**	0.70	0.52	**0.23**

Gender+ = stereotype-inducing prompts ("successful CEO").
One "person" vector $\rightarrow$ all professions; UCE retrains per-profession.

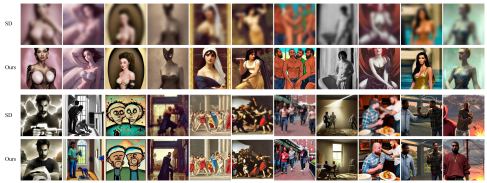
SD

Ours

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Complementary to Existing Safety Methods

h-space steering is orthogonal to both fine-tuning and guidance, so it stacks on top of existing safety methods.



I2P score ($\downarrow$ = safer)

Method	I2P	Δ
SD	0.41	—
+ **Ours**	**0.27**	−34%
SLD	0.20	—
+ **Ours**	**0.16**	−20%
ESD	0.32	—
+ **Ours**	**0.27**	−16%

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Quality Preserved

Safety interventions often degrade generation quality. Ours does not.

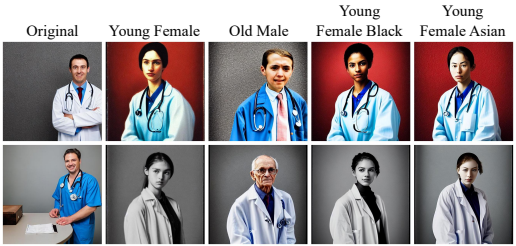
COCO-3K (FID ↓ / CLIP ↑)

Method	FID ↓	CLIP ↑
SD	**14.09**	31.33
SLD	19.35	30.41
ESD	15.36	30.05
Ours	15.98	**31.03**

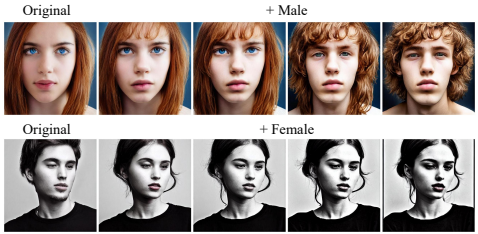
Other methods sacrifice fidelity (SLD: FID +5.3) or alignment (ESD: CLIP −1.3), while *h*-space steering achieves minimal degradation with the best text-image alignment.

Composition & Interpolation

Nice properties of the semantic space and concept vectors:
learned vectors compose and interpolate cleanly.



Independently learned vectors (gender, age, race) combine linearly without interference.



Varying $\|\Delta h\|$ produces smooth transitions that are disentangled from other factors.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Domain Transfer

The "running" vector learned from dogs applies to cats and humans, showing that it captures a concept rather than a pixel pattern.

The prompt "elephant wearing glasses" fails on its own, but adding $\Delta h_{\text{glasses}}$ renders them at the correct location.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Takeaway

Contributions

- Self-supervised concept discovery in h-space that requires no external classifier, no human labels, and applies to any concept
- Two responsible-generation applications as inference-time vector addition: fair sampling and safe suppression

Limitations

- A single Δh is shared across all timesteps, but per-step vectors may capture finer dynamics
- Requires prompt pairs (y^+/y^-); not all concepts are easily isolable in text

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

From Interpretability to Control

The interpretability artifact is itself the intervention.

- Locating a concept in *h*-space does not merely explain a failure; the same vector is the mechanism that repairs it.
- The intervention is training-free: no fine-tuning, no classifier, no per-concept data collection — and it composes with methods that do retrain.
- Reading a model's internal representation is therefore a practical route to *controllable* generation, not only a diagnostic one.

Where this goes next

- Extend *h*-space self-discovery to spatio-temporal latents in DiT-based video models, enabling steering of unsafe content across time
- Per-timestep directions, to capture how a concept is formed during denoising

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.

Thank You

Questions?

Self-Discovering Interpretable Diffusion Latent Directions
for Responsible Text-to-Image Generation

Hang Li · Chengzhi Shen · Philip Torr · Volker Tresp · Jindong Gu

CVPR 2024

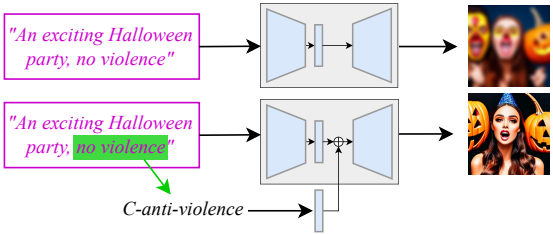
https://interpretdiffusion.github.io/

Backup: Responsible Text-Enhancing Generation

A third application: extract safety-related concepts *from the prompt itself* and activate the corresponding learned vectors.

$$\mathrm{h} \leftarrow \mathrm{h} + \mathrm{c}(y), \qquad \mathrm{c}(y) = \sum_i \Delta \mathrm{h}_i$$

- Models often fail to honour responsible phrasing — e.g. "no violence", because negation is poorly handled by the text encoder.
- The learned vector reinforces the desired ethical concept during generation, rather than relying on the prompt alone.
- Same mechanism as fair and safe generation: inference-time addition in *h*-space.



Quantitative results for this setting are in the paper.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation, CVPR 2024.