Controlling Generative Models through Parameter Localization

Łukasz Staniszewski

ML in PL Conference 2025

“We must develop the ability to sustainably adapt to unexpected challenges in AI. ***Resilience*** *is about how the ecosystem responds when something completely unexpected happens.”*

~David Bau

Assistant Professor Northeastern Khoury College
Leader of Bau Lab

*“We must develop the ability to sustainably adapt to unexpected challenges in AI. **Resilience** is about how the ecosystem responds when something completely unexpected happens.”*

*“To develop a resilient AI ecosystem, we must invest in: the science of **understanding AI**, the technical practice of **control of AI**, and a culture of **power over AI**.”*

~David Bau

Assistant Professor Northeastern Khoury College
Leader of Bau Lab

Controlling Generative Models through Parameter Localization

Łukasz Staniszewski

Controlling Generative Models through Parameter Localization

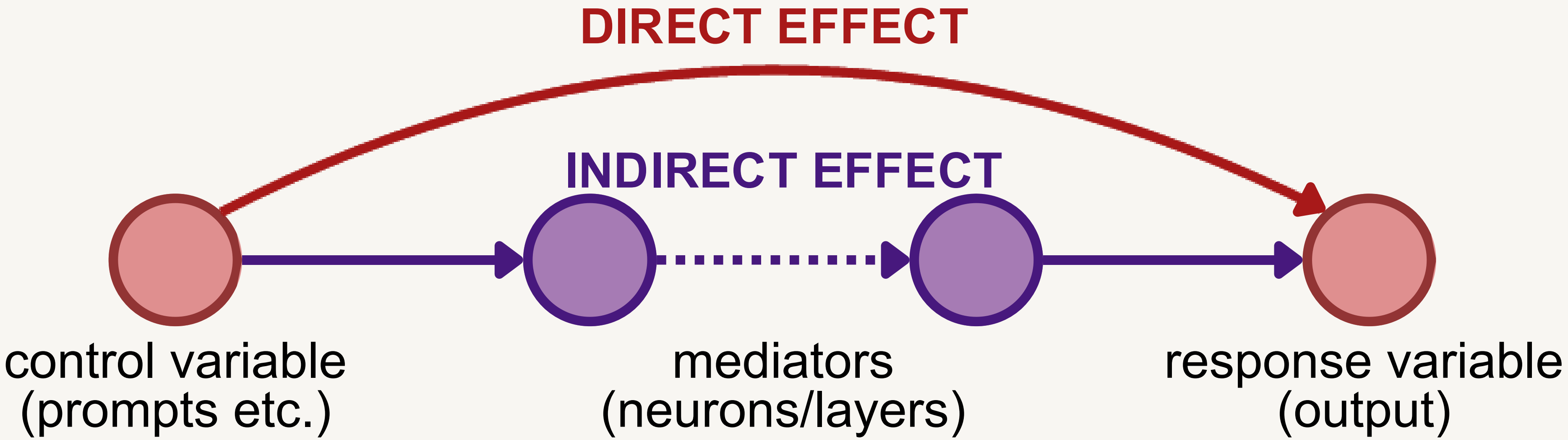
Łukasz Staniszewski

Thanks to (alphabetical order):

- **Franziska Boenisch**
- **Bartosz Cywiński**
- **Kamil Deja**
- **Adam Dziedzic**
- **Mateusz Modrzejewski**
- **Katarzyna Zaleska**

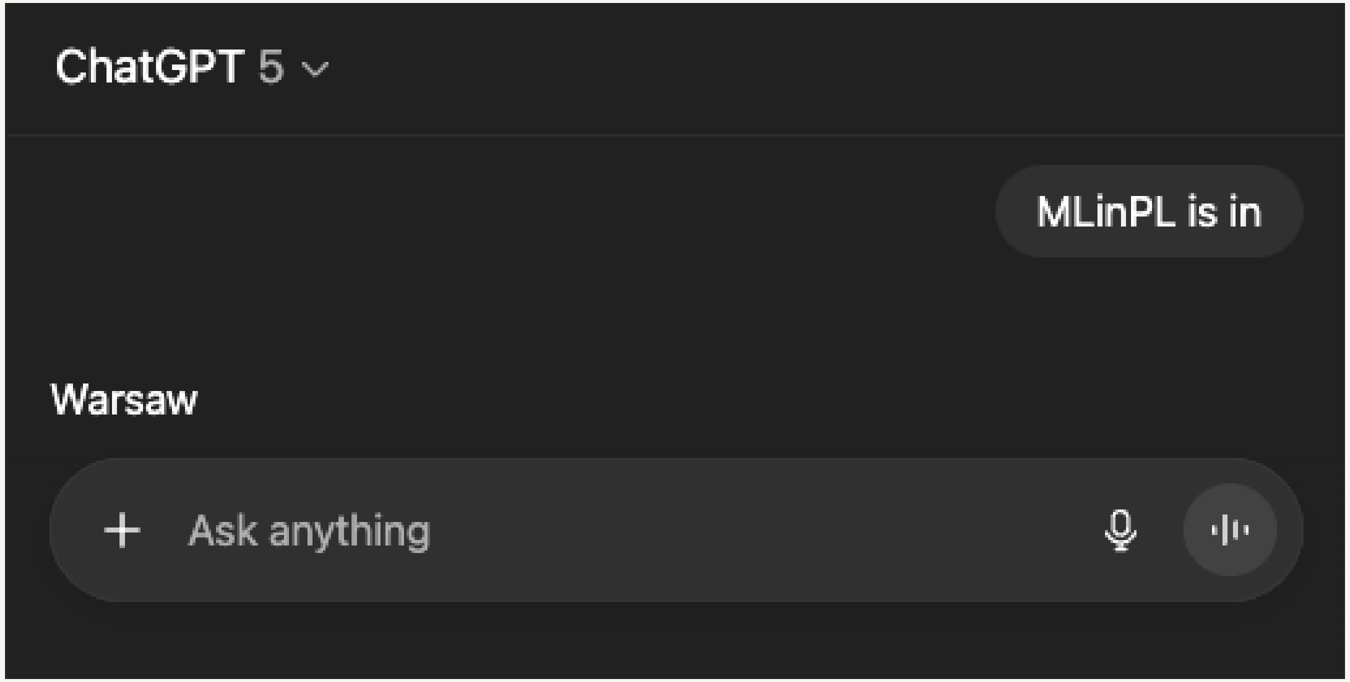
What is Localization about?

Causal Mediation Analysis [1] studies the change in a response variable following an active intervention on intermediate variables of interest.

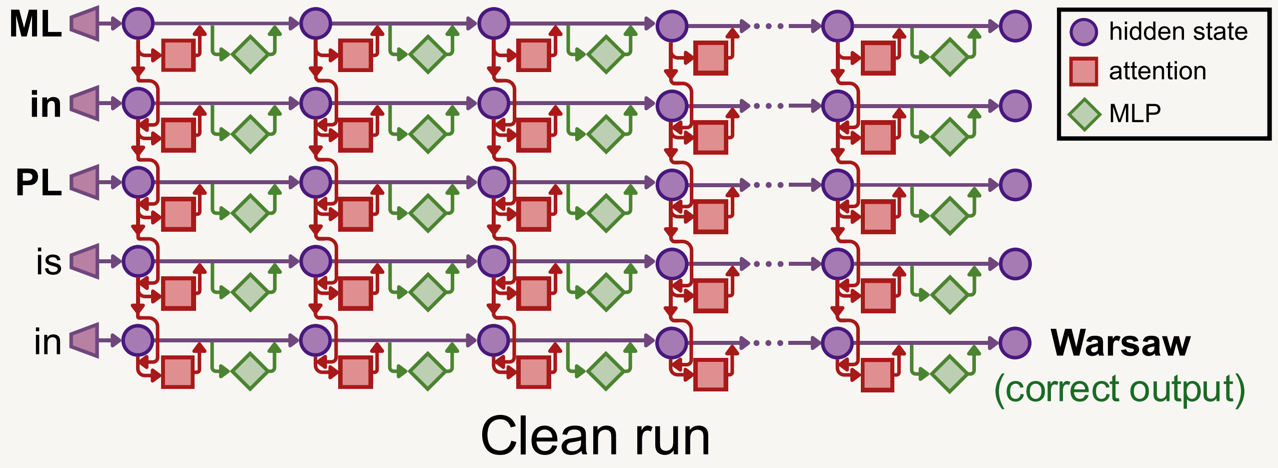


[1] Pearl, Judea. "Direct and Indirect Effects." Probabilistic and Causal Inference: The Works of Judea Pearl (2001).

What is Localization about?

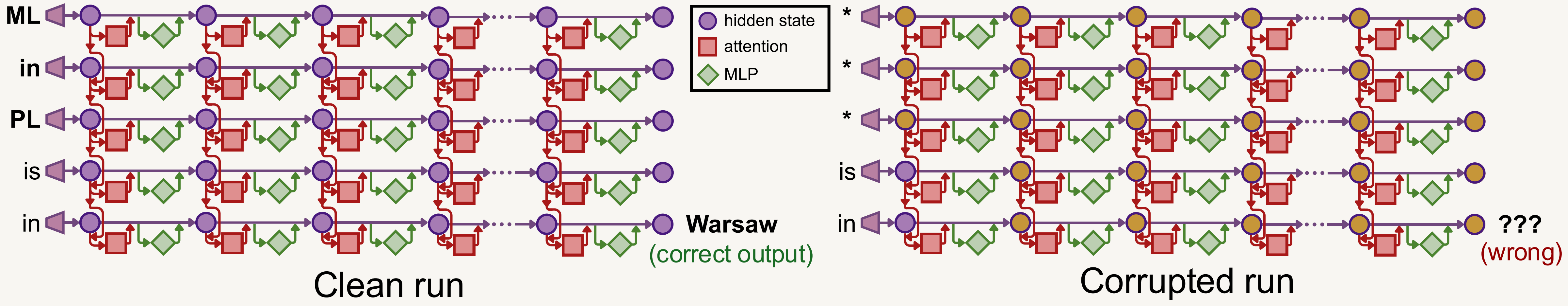


What is Localization about?



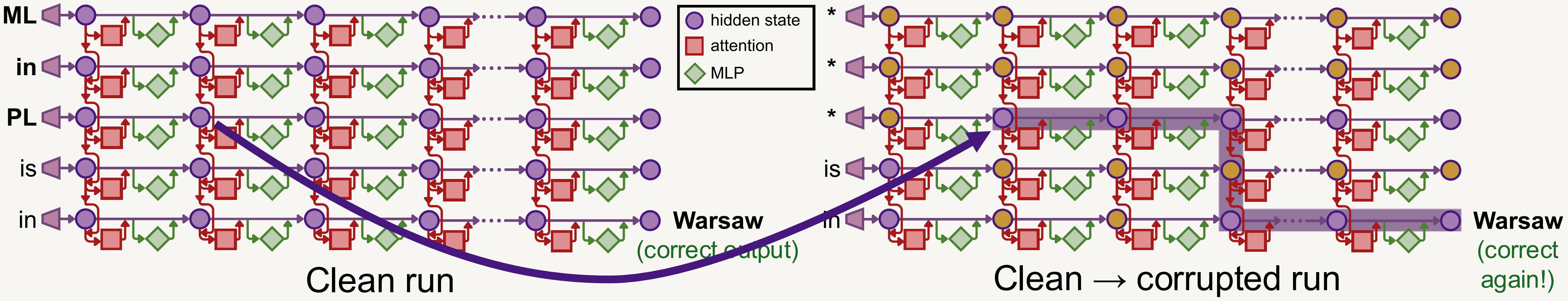
[2] Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

What is Localization about?



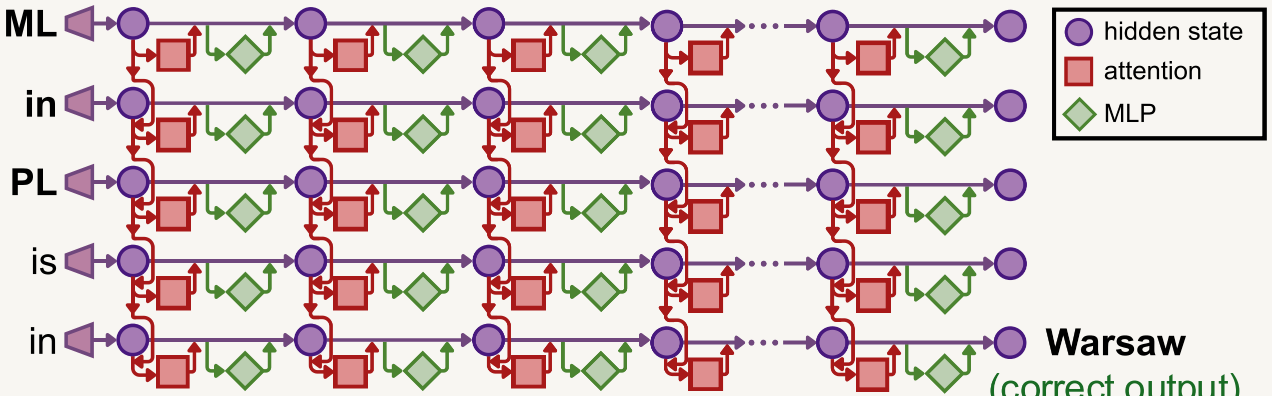
[2] Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

What is Localization about?



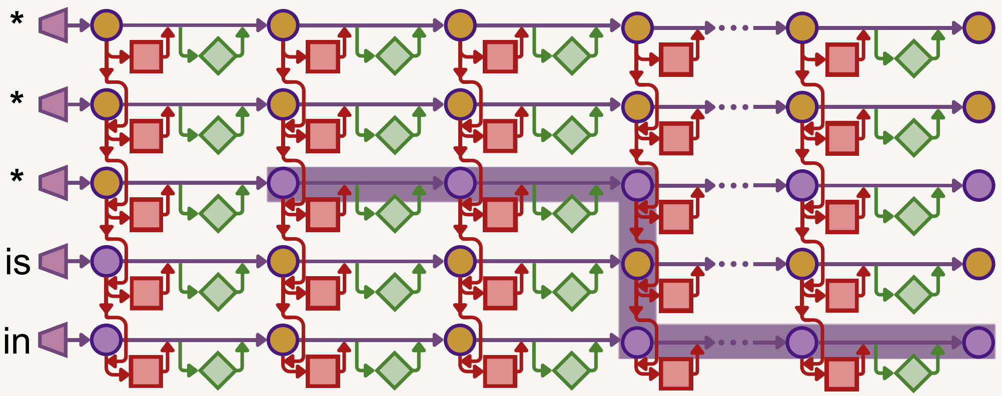
[2] Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

What is Localization about?



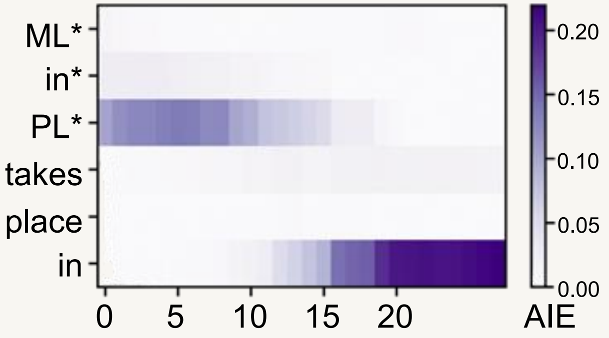
Warsaw (correct output)

Clean run

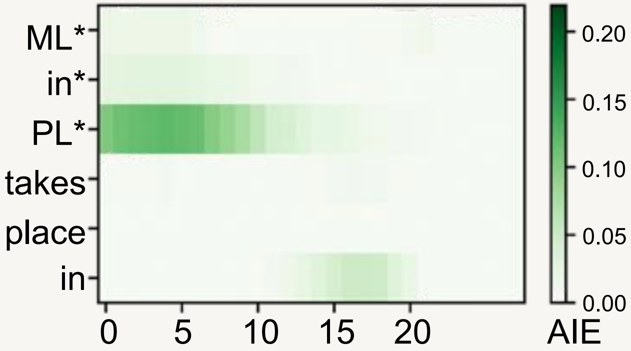


Warsaw (correct again!)

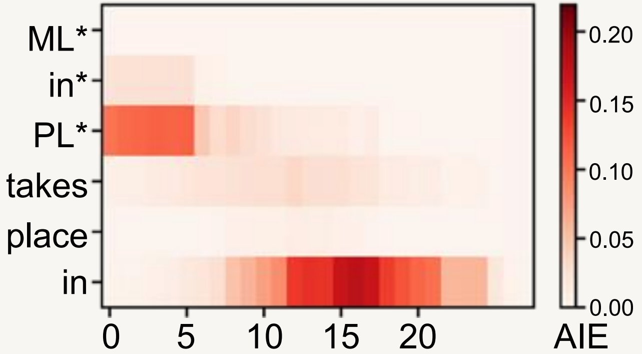
Clean → corrupted run



Patching **hidden states**



Patching **MLPs**



Patching **Attention**

[2] Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

What is Localization about?

“Localized Factual Association Hypothesis” in LLMs:

- midlayer MLPs → key-value mappers that recall memorized properties (Warsaw) for the subjects (MLinPL) - **knowing**;
- attention layers → copy the information accumulated by MLPs to the last token - **saying**.

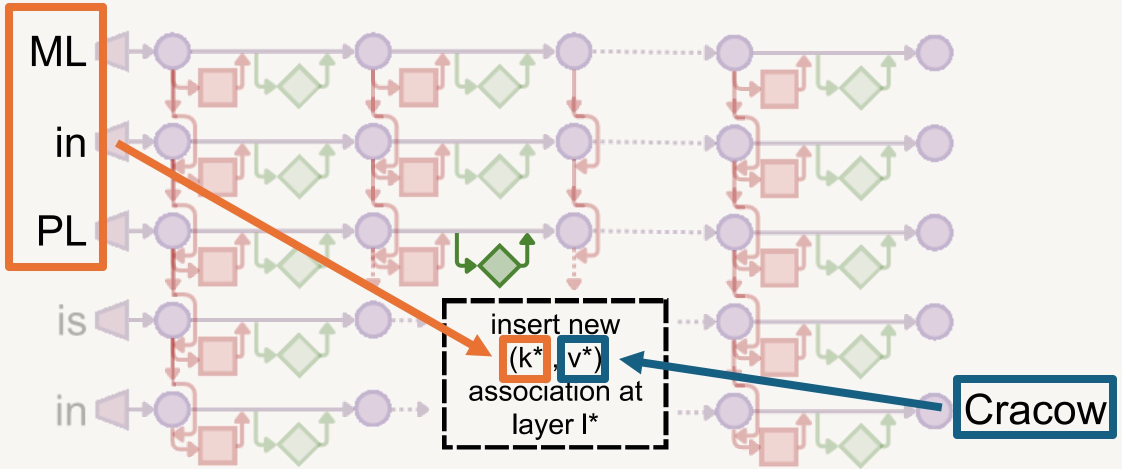
[2] Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

What is Localization about?

“Localized Factual Association Hypothesis” in LLMs:

- midlayer MLPs → key-value mappers that recall memorized properties (Warsaw) for the subjects (MLinPL);
- attention layers → copy the information accumulated by MLPs to the last token.

ROME [2] - Rank-One Model Editing:



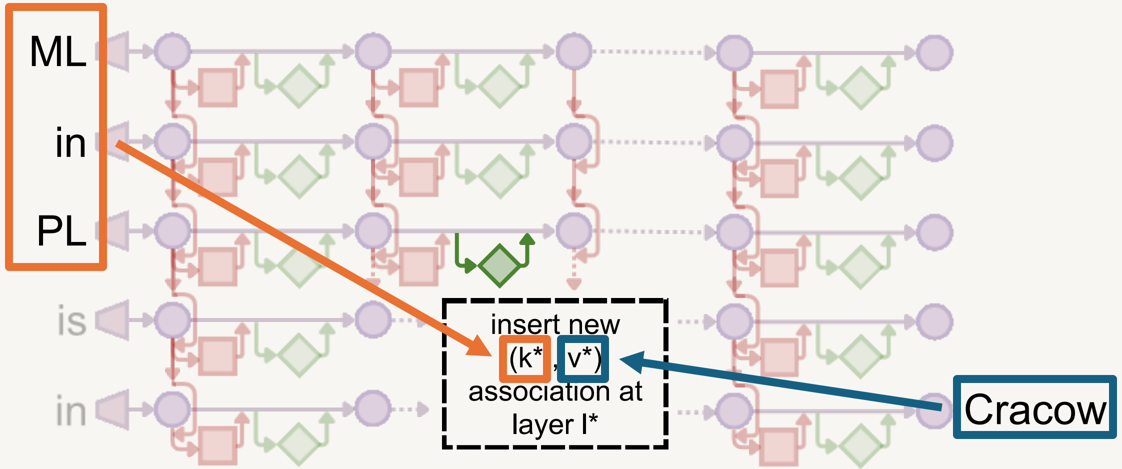
[2] Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

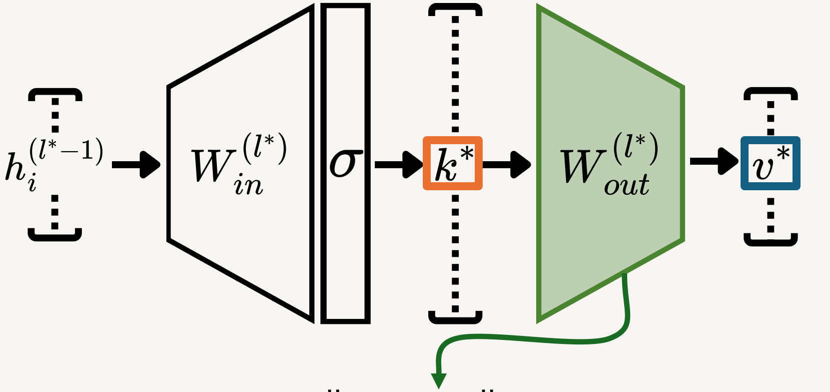
What is Localization about?

“Localized Factual Association Hypothesis” in LLMs:

- midlayer MLPs $\rightarrow$ key-value mappers that recall memorized properties (Warsaw) for the subjects (MLinPL);
- attention layers $\rightarrow$ copy the information accumulated by MLPs to the last token.

ROME [2] - Rank-One Model Editing:

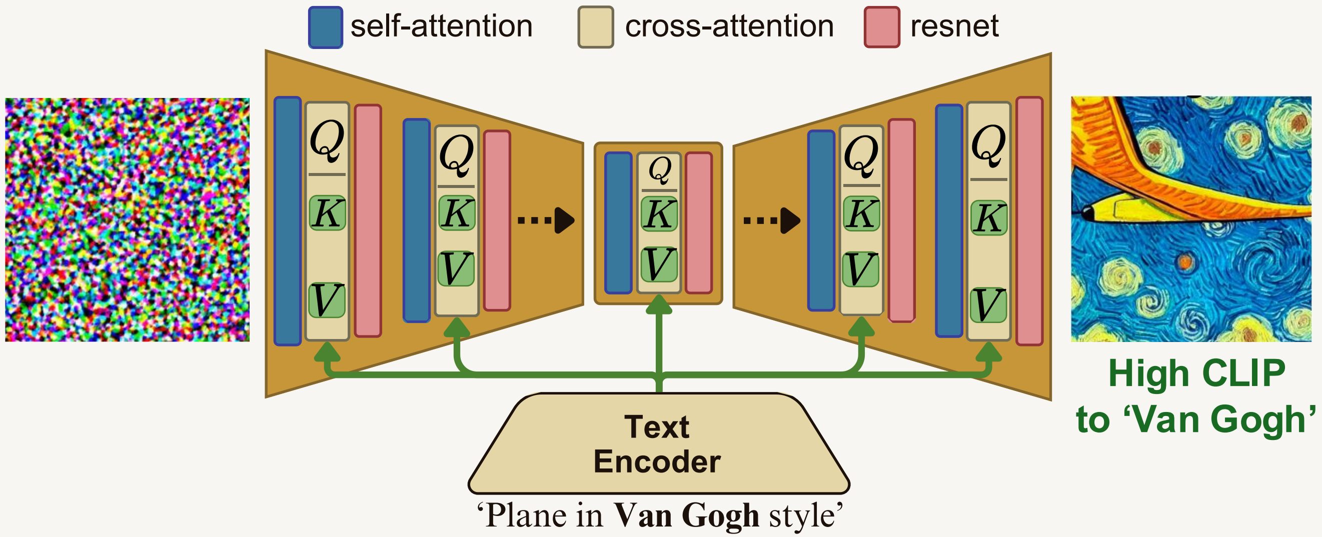




$\widehat{\boldsymbol{W}} \coloneqq$ minimize $\|\widehat{\boldsymbol{W}}\boldsymbol{K} - \boldsymbol{V}\|$ such that $\widehat{\boldsymbol{W}}\boldsymbol{k}^* = \boldsymbol{v}^*$

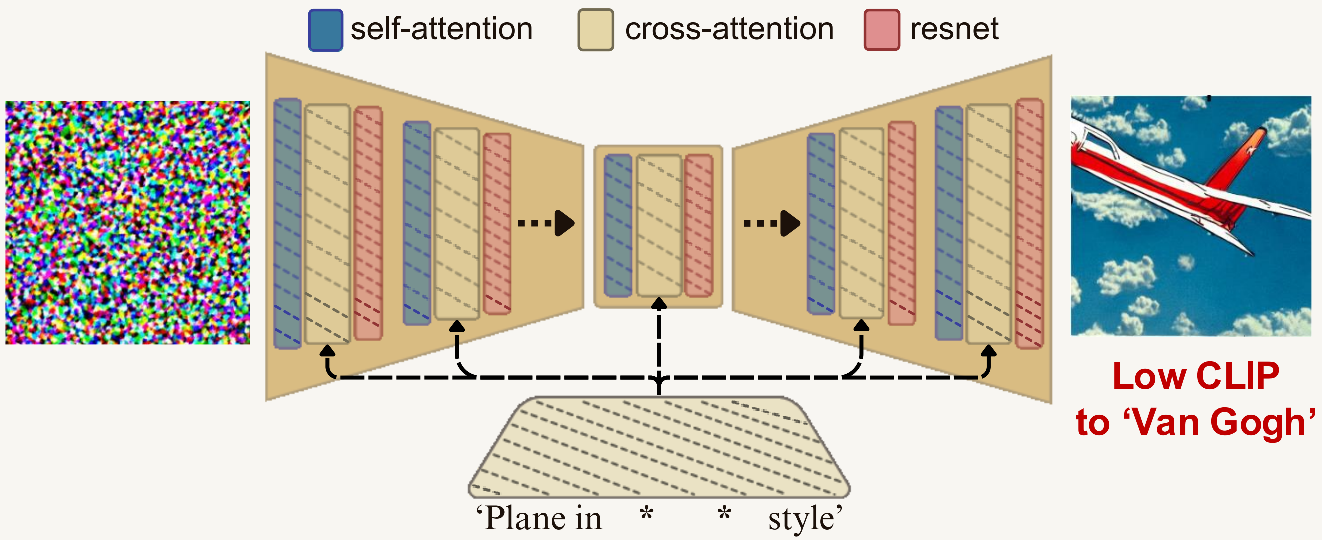
[2] Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

Localization in Text-to-Image models



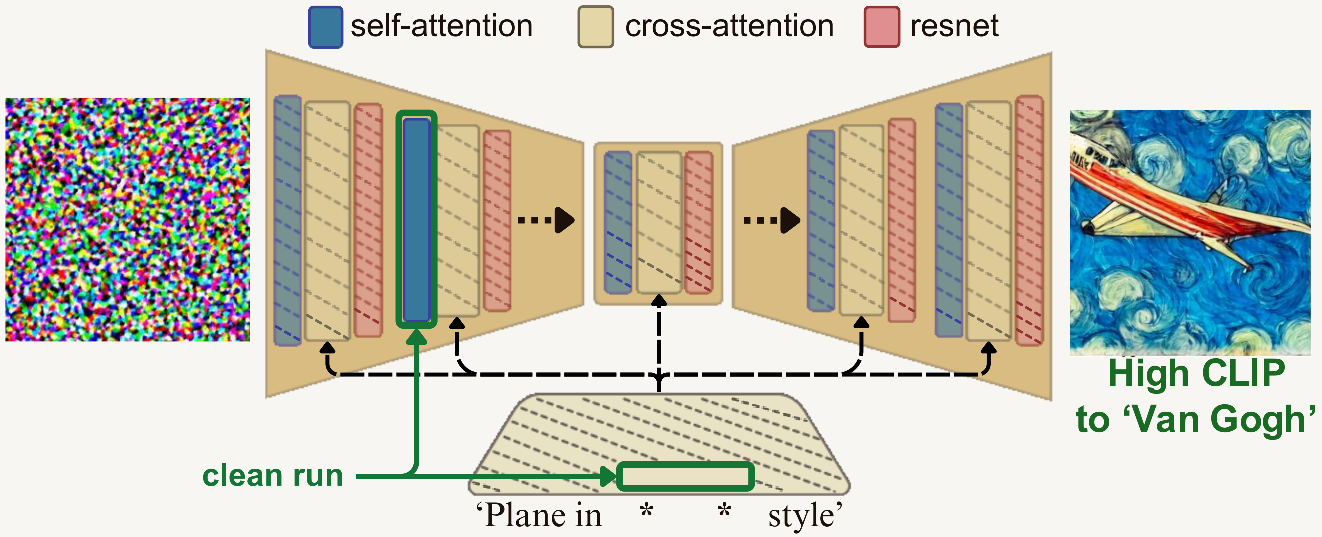
[3] Basu, Samyadeep, et al. "Localizing and editing knowledge in text-to-image generative models." ICLR 2023.

Localization in Text-to-Image models



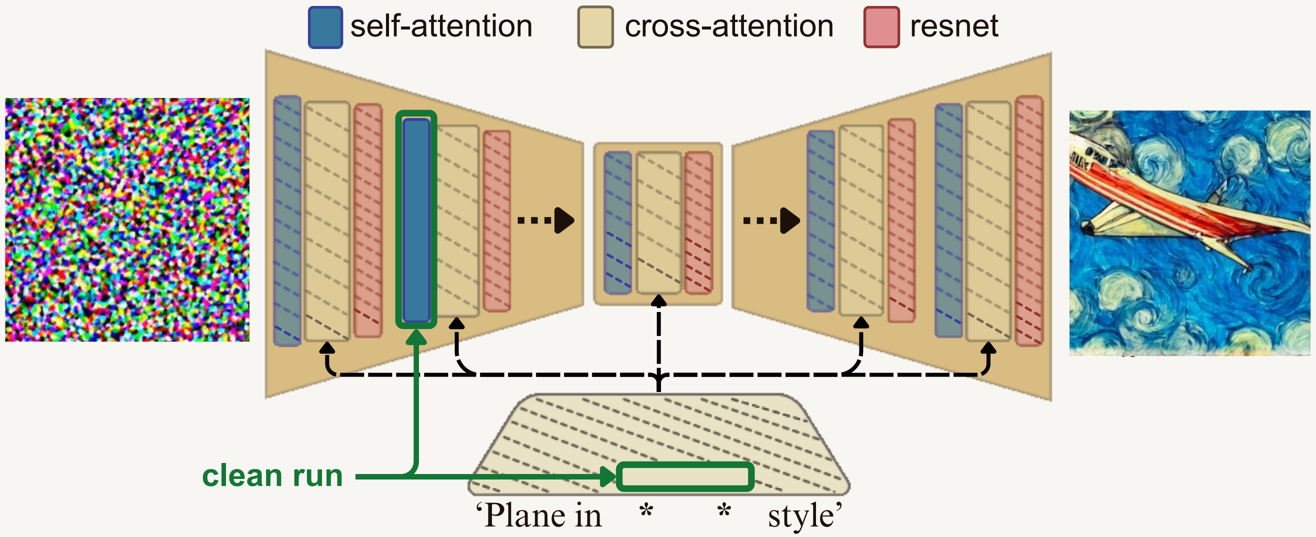
[3] Basu, Samyadeep, et al. "Localizing and editing knowledge in text-to-image generative models." ICLR 2023.

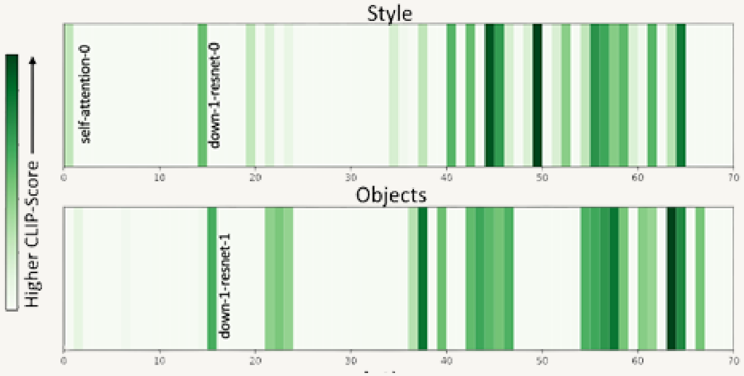
Localization in Text-to-Image models



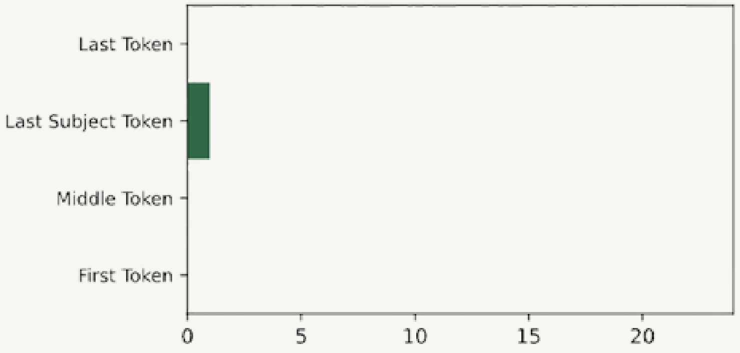
[3] Basu, Samyadeep, et al. "Localizing and editing knowledge in text-to-image generative models." ICLR 2023.

Localization in Text-to-Image models





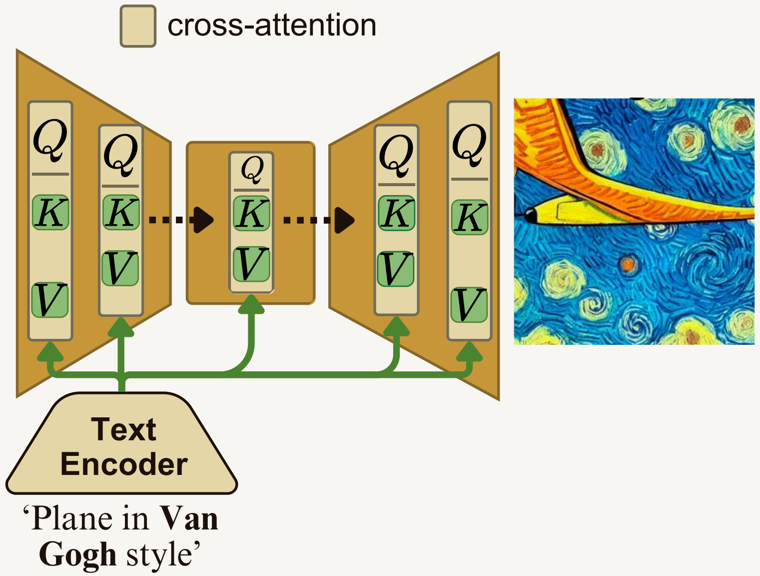
U-Net layers



text encoder layers

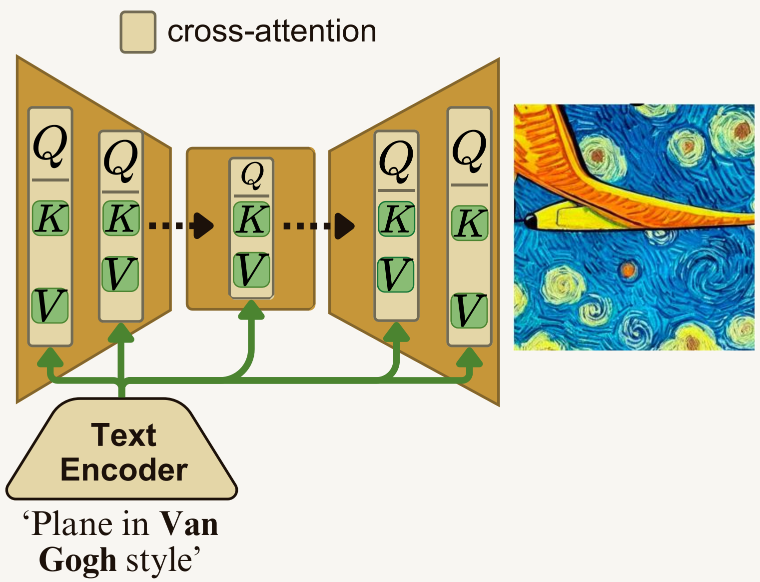
[3] Basu, Samyadeep, et al. "Localizing and editing knowledge in text-to-image generative models." ICLR 2023.

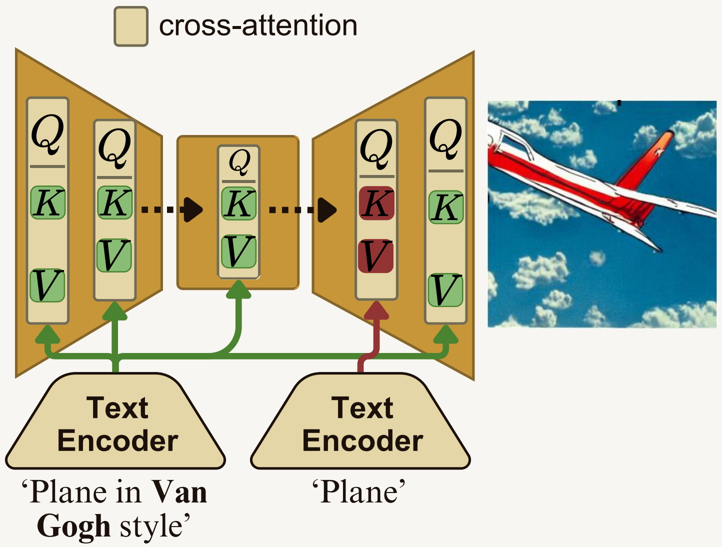
Localization in Text-to-Image models



[4] Basu, Samyadeep, et al. "On mechanistic knowledge localization in text-to-image generative models." ICML 2024.

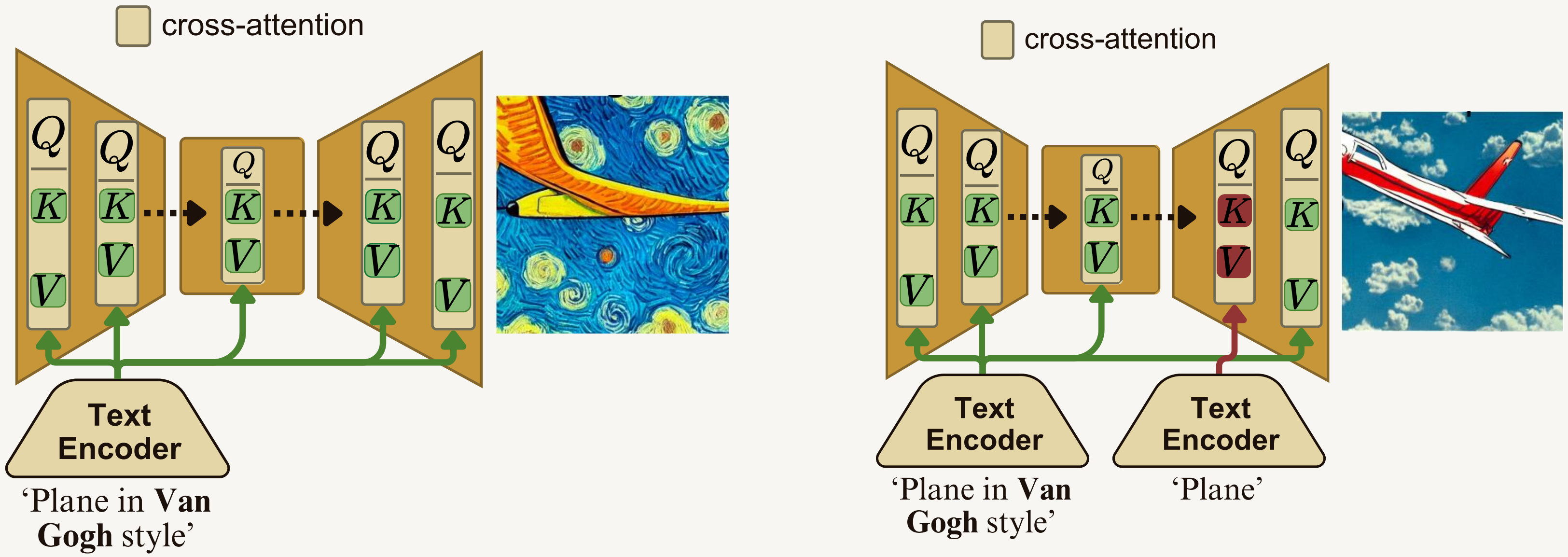
Localization in Text-to-Image models

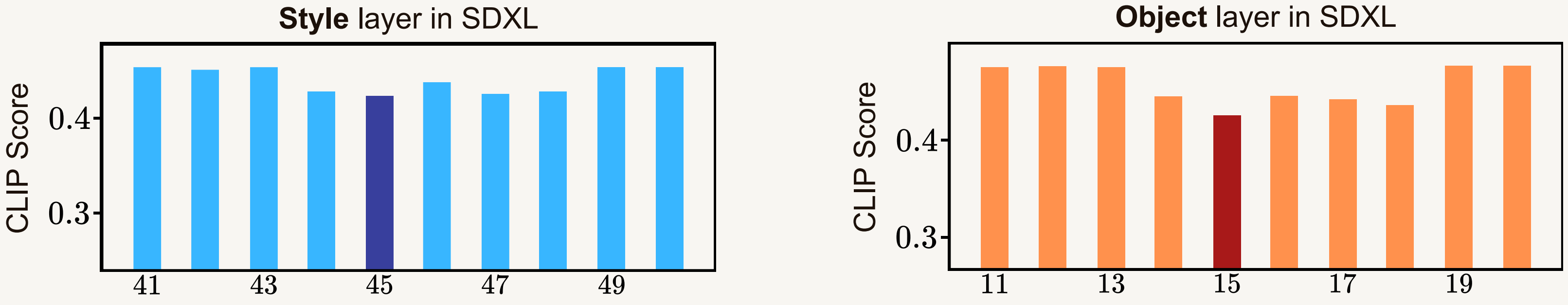




[4] Basu, Samyadeep, et al. "On mechanistic knowledge localization in text-to-image generative models." ICML 2024.

Localization in Text-to-Image models





[4] Basu, Samyadeep, et al. "On mechanistic knowledge localization in text-to-image generative models." ICML 2024.

Text-to-image models are great at generating text!

Stable Diffusion 3 [4]

[4] Esser, Patrick, et al. "Scaling rectified flow transformers for high-resolution image synthesis." ICML 2024.

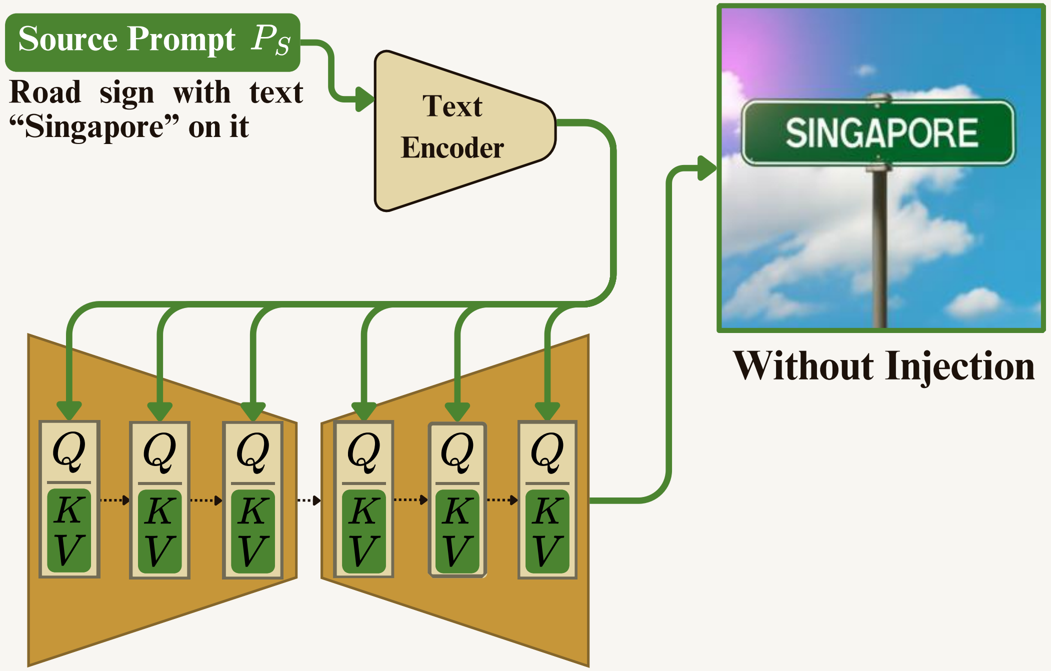
Text-to-image models are too great at generating text!

Stable Diffusion 3 [4]

FLUX.1 [5]

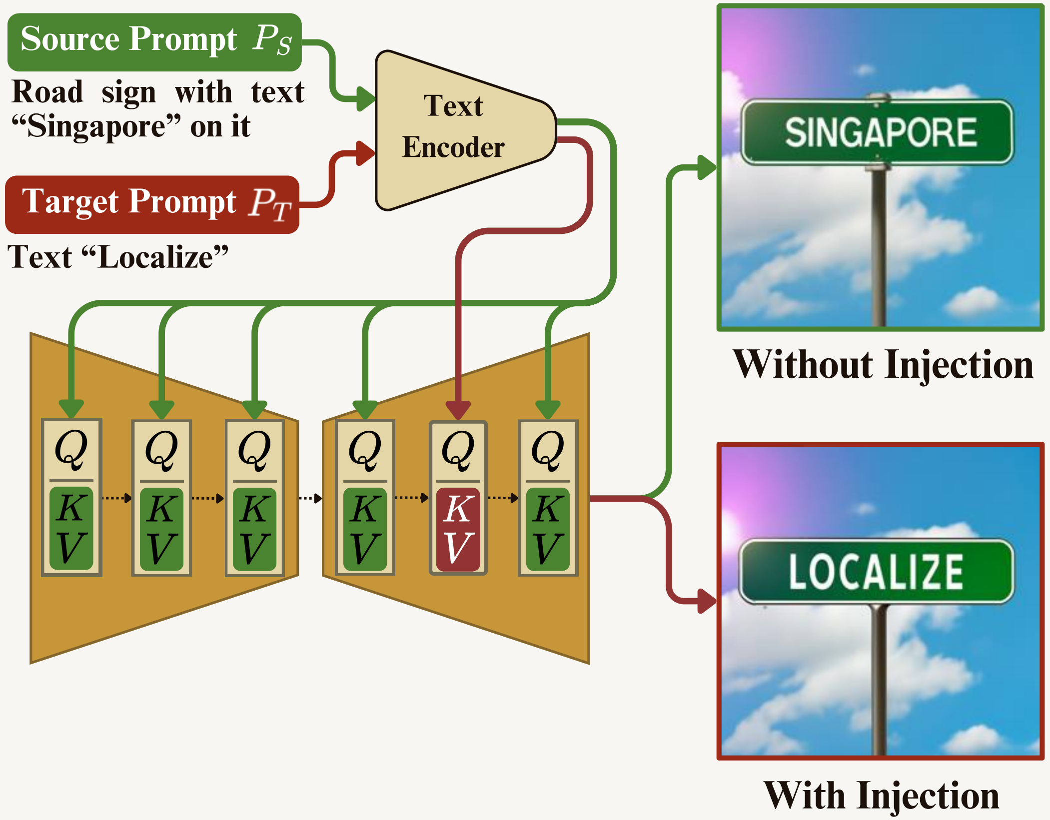
[4] Esser, Patrick, et al. "Scaling rectified flow transformers for high-resolution image synthesis." ICML 2024.
[5] Black Forest Labs. Flux.1, 2024.

Controlling Text Content in U-Nets



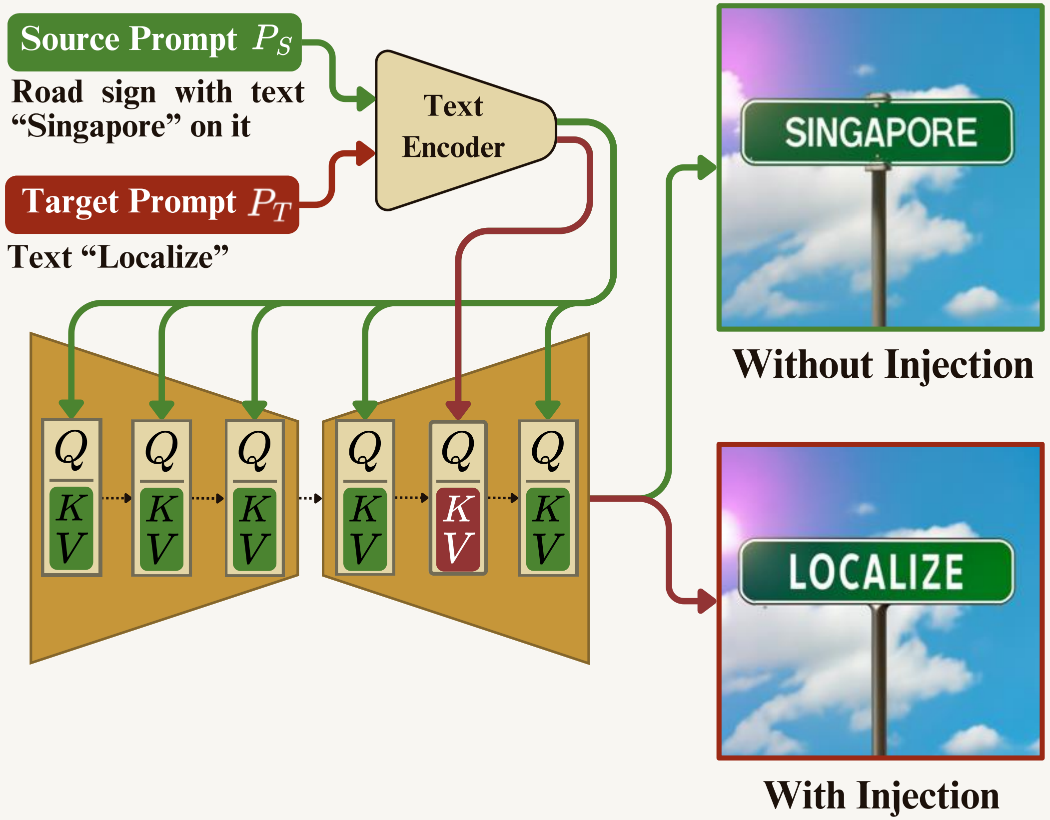
[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

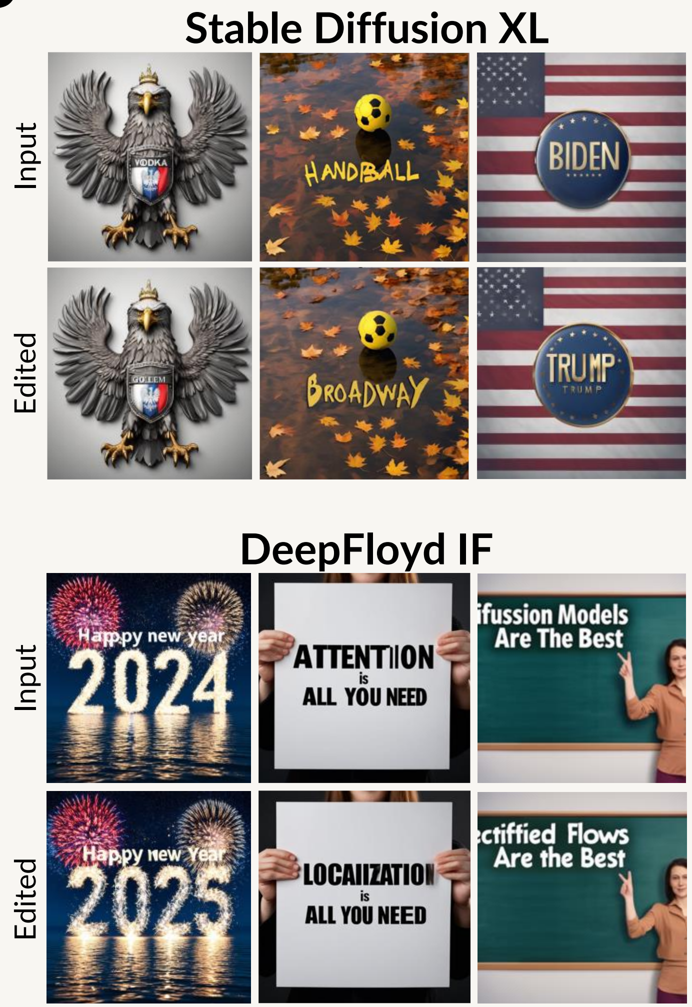
Controlling Text Content in U-Nets



[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

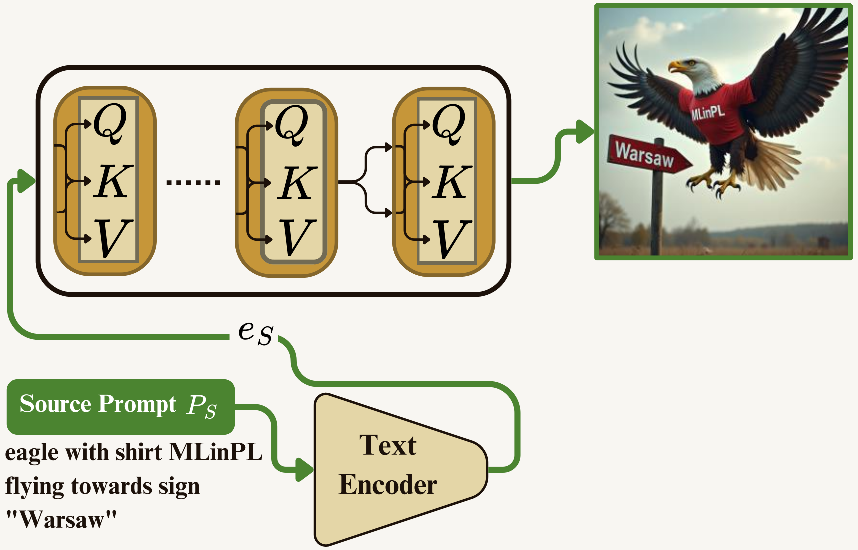
Controlling Text Content in U-Nets





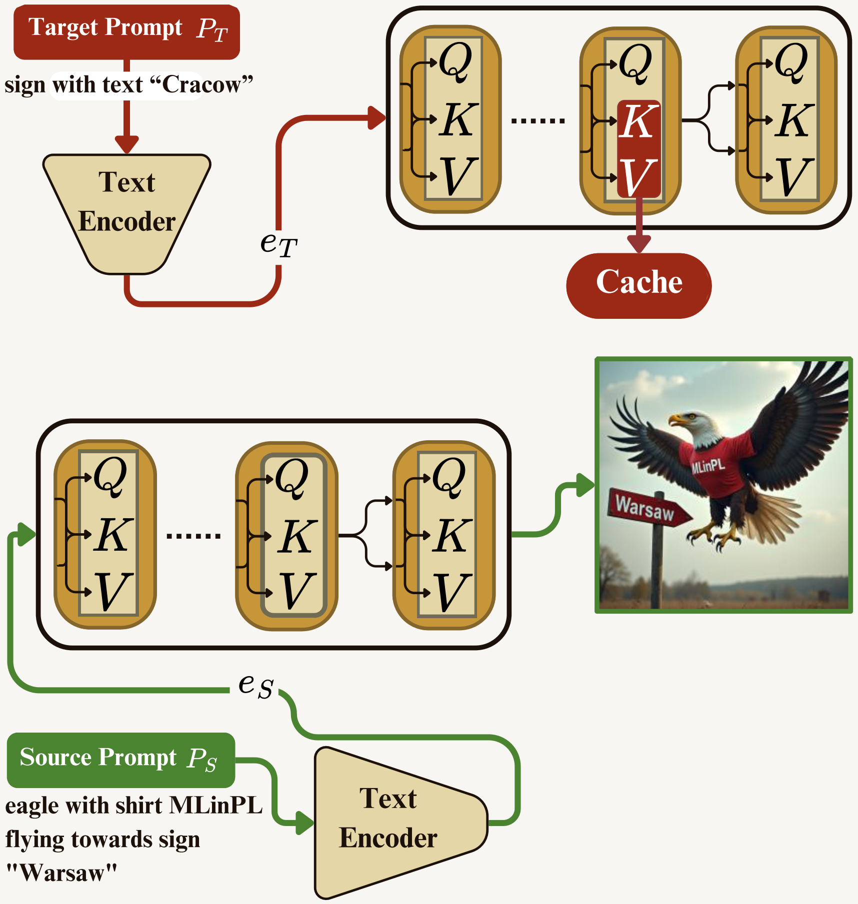
[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content in Diffusion Transformers



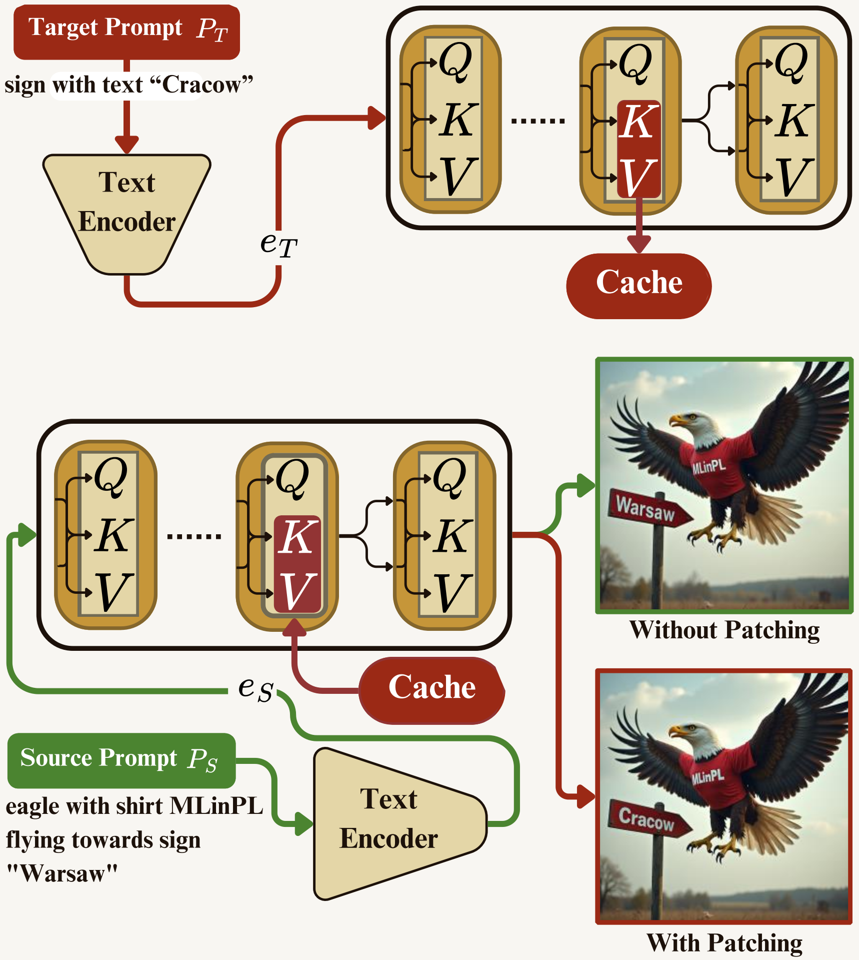
[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content in Diffusion Transformers



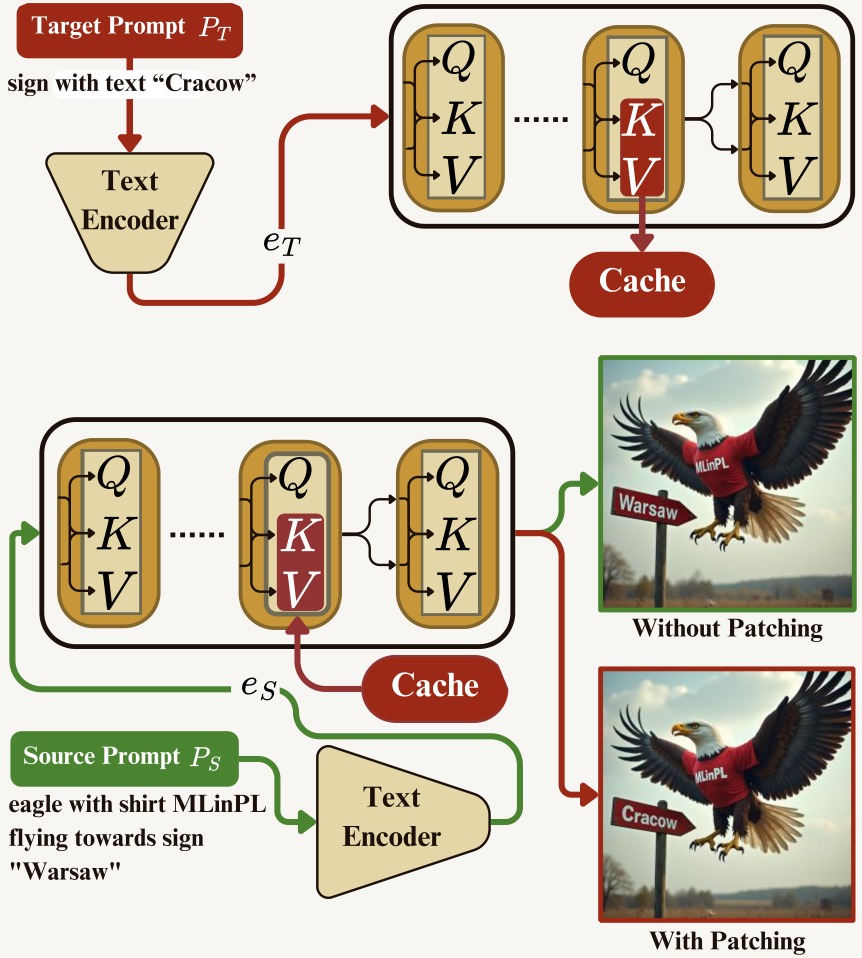
[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models.” ICLR 2025.

Controlling Text Content in Diffusion Transformers



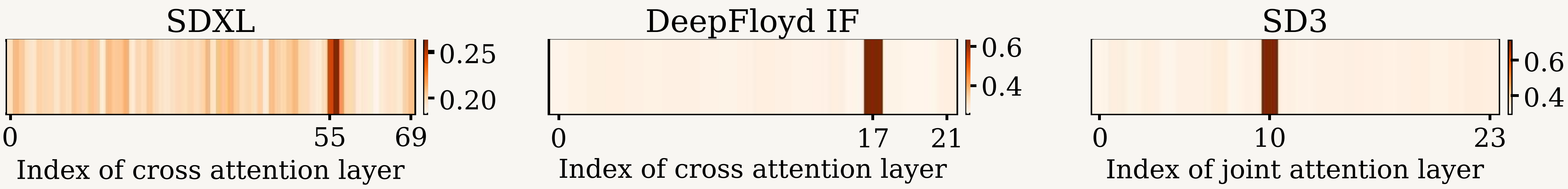
[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content in Diffusion Transformers



[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content – localization results



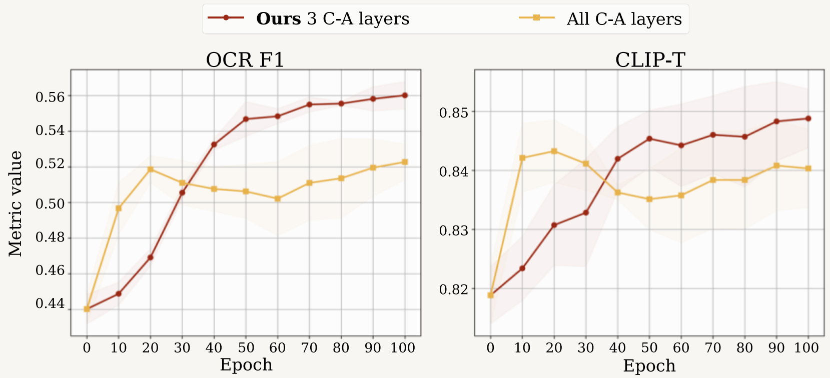
Model name	# localized cross attention layers	# total cross attention layers	% of model parameters
Stable Diffusion XL	3	70	0.61%
DeepFloyd IF	1	22	0.21%
Stable Diffusion 3	1	24	0.23%

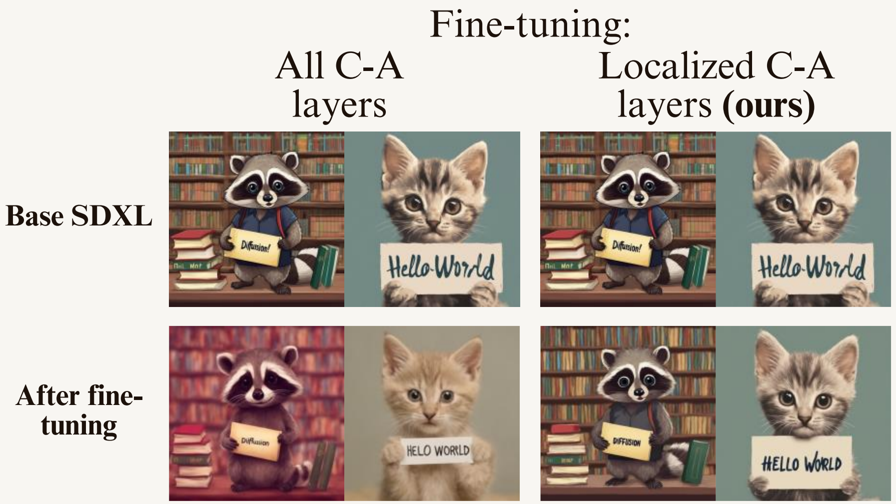
Table 1: Localized layers in numbers.

[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content – applications?

Our **localization-based fine-tuning strategy**, which targets only the localized layers, improves text generation and maintains generation diversity.

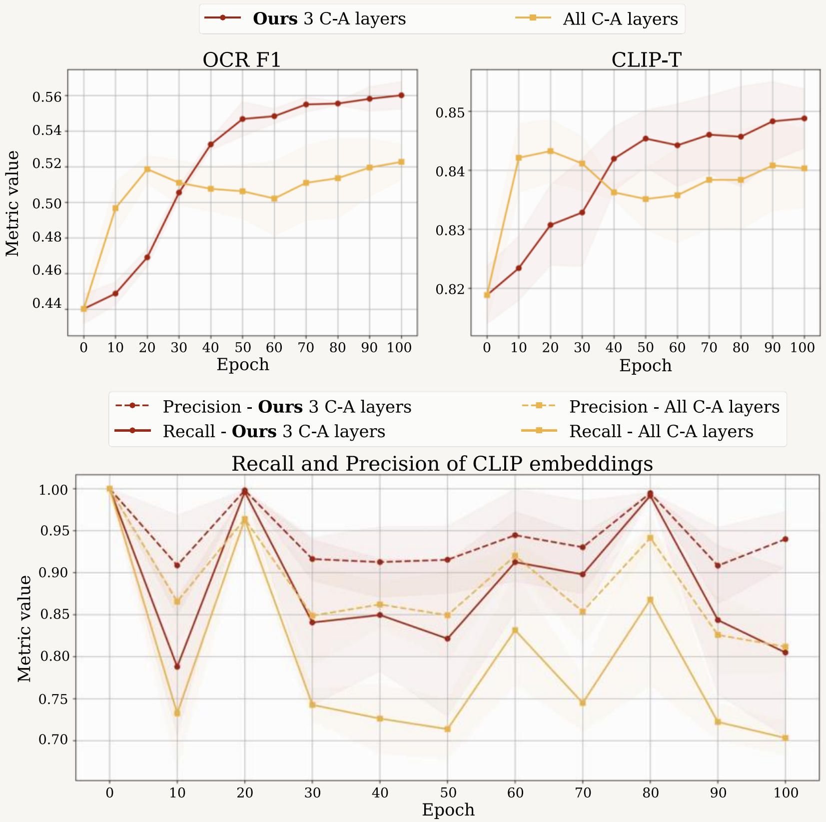


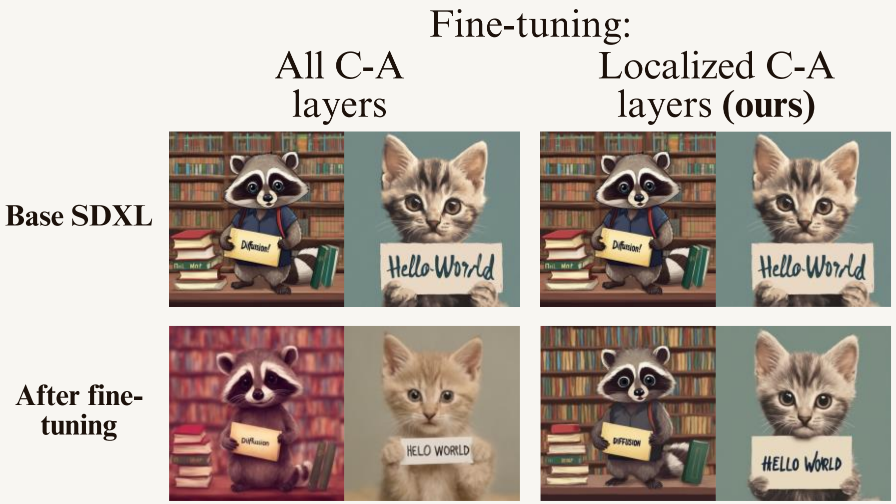


[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content – applications?

Our **localization-based fine-tuning strategy**, which targets only the localized layers, improves text generation and maintains generation diversity.





[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content – applications?

Our text-editing method successfully prevents the **generation of a toxic text** within images in just one forward pass while maintaining the background.

Method	**Model**	SSIM ↑	OCR F1 ↓	Toxicity score ↓
Negative prompt	SDXL	0.71	0.23	0.052
Safe Diffusion*	SDXL	**0.81**	0.33	0.209
Ours	SDXL	0.79	**0.20**	**0.003**
Negative prompt	IF	0.37	0.59	0.250
Safe Diffusion*	IF	**0.74**	0.79	0.540
Ours	IF	0.61	<u>0.32</u>	**0.018**
Negative prompt	SD3	0.53	0.77	0.407
Safe Diffusion*	SD3	**0.87**	0.73	0.568
Ours	SD3	0.70	**0.32**	**0.018**

Original Image | Negative Prompt | Safe Diffusion | **Ours**

[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Controlling Text Content - limitations

- Localized layers control textual content → we **cannot**:
  - add text to the image;
  - change position of the text;
  - change style of the text.

[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

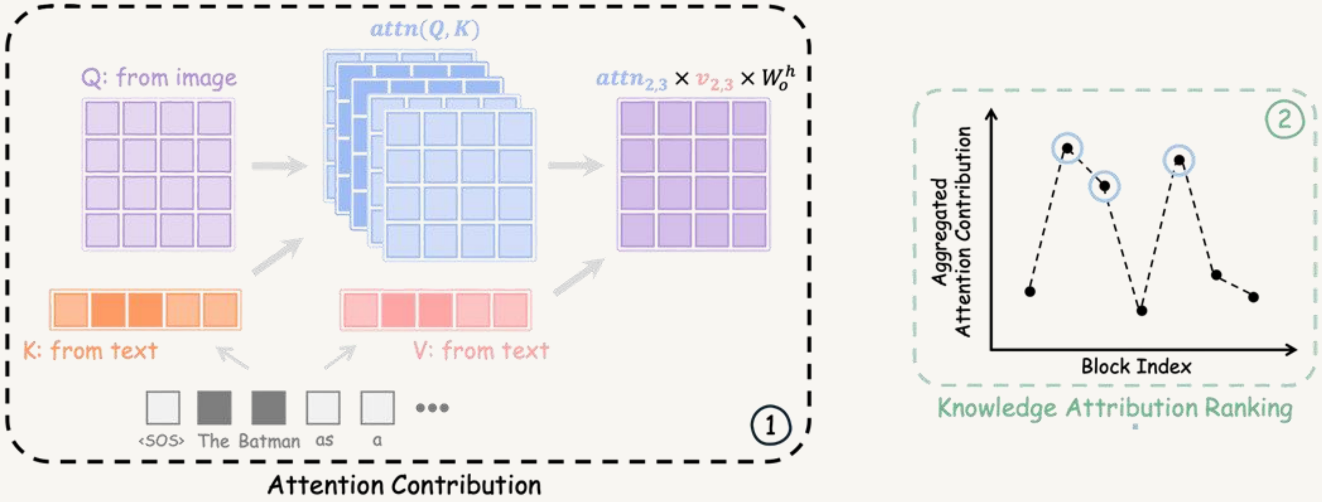
Controlling Text - limitations

- Localized layers control textual content → we **cannot**:
  - add text to the image;
  - change position of the text;
  - change style of the text.

- The localization process takes some time → to find important layers among 50 Cross Attentions, we need to do 1+50 inferences.

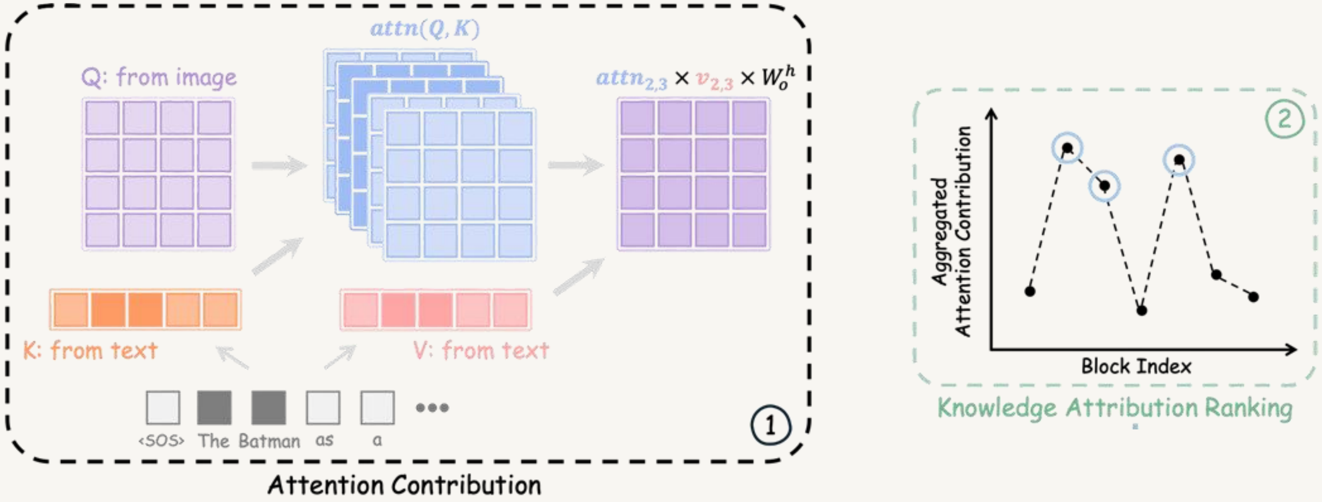
[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

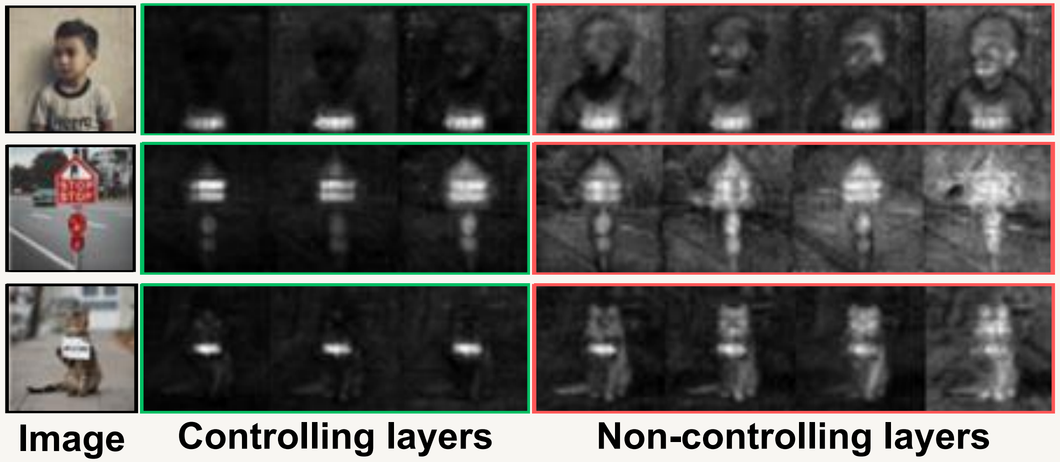
Localization – can we make it faster?



[7] Zarei, Arman, et al. "Localizing Knowledge in Diffusion Transformers." arXiv preprint arXiv:2505.18832 (2025).

Localization – can we make it faster?

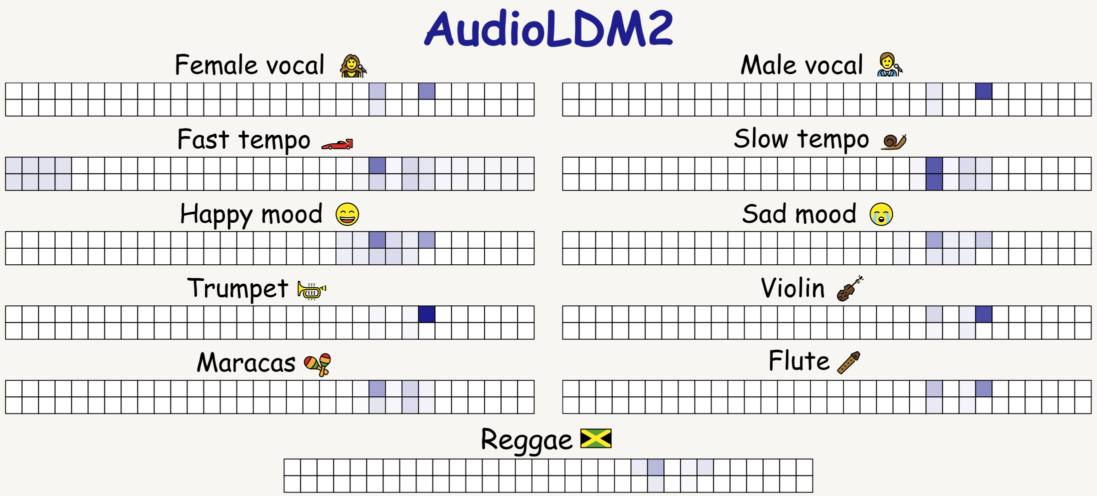


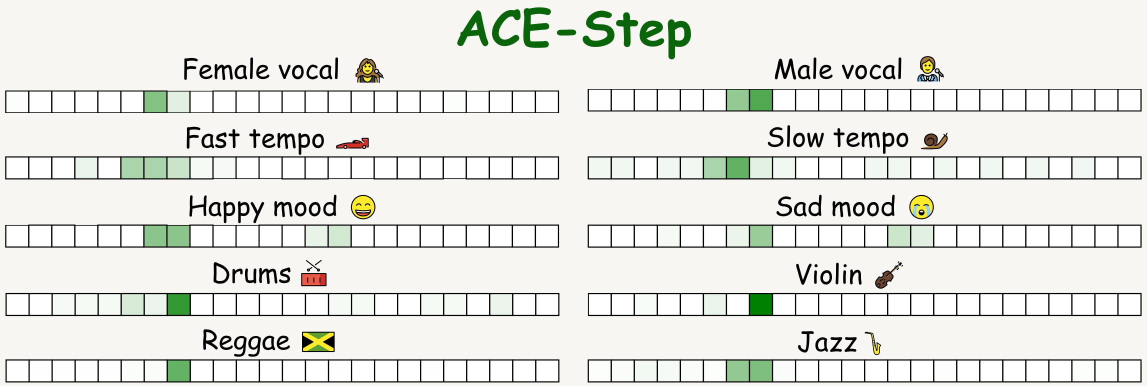


[7] Zarei, Arman, et al. "Localizing Knowledge in Diffusion Transformers." arXiv preprint arXiv:2505.18832 (2025).

How about audio generation models?

CLAP similarity to musical concepts in AudioLDM2 and ACE-Step models. The control over diverse audio concepts is concentrated and shared by single layers.





Conclusions

- Making AI ecosystems resilient is crucial and can be achieved by localizing key model components.
- Localization is possible across text, image, audio (and possibly other) generation models.
- Localized layers can be used effectively for many downstream tasks:
  - model knowledge: editing, unlearning, personalization;
  - image editing;
  - robust fine-tuning;
  - removing toxic content from generations.

Thanks!

luks.staniszewski@gmail.com

Based on:

Pearl, Judea. "Direct and Indirect Effects." Probabilistic and Causal Inference: The Works of Judea Pearl (2001).

Meng, Kevin, et al. "Locating and Editing Factual Associations in GPT." NeurIPS 35 (2022).

Basu, Samyadeep, et al. "Localizing and editing knowledge in text-to-image generative models." ICLR 2023.

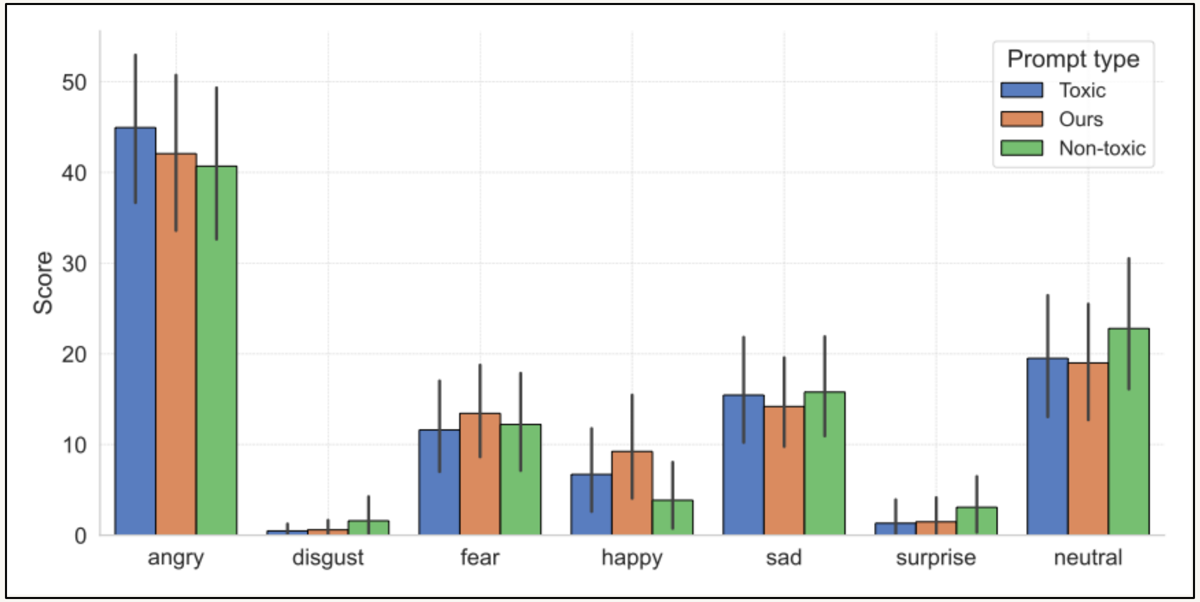
Basu, Samyadeep, et al. "On mechanistic knowledge localization in text-to-image generative models." ICML 2024.

Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Zarei, Arman, et al. "Localizing Knowledge in Diffusion Transformers." arXiv preprint arXiv:2505.18832 (2025).

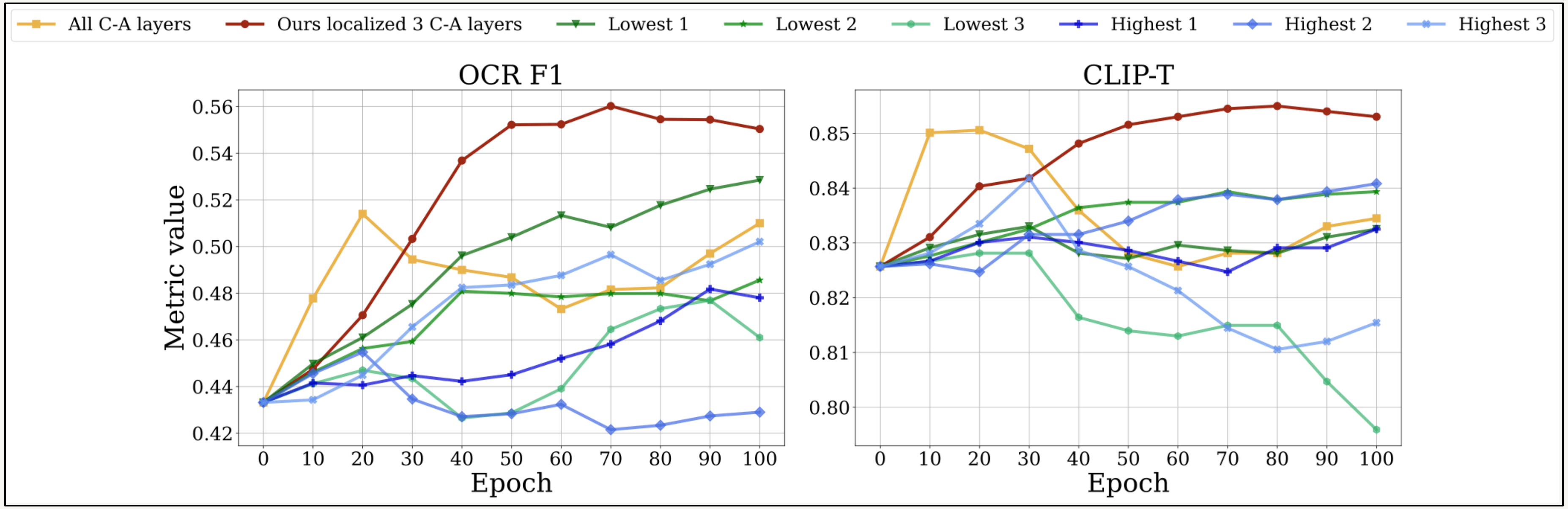
Backup slides

Swapping prompts



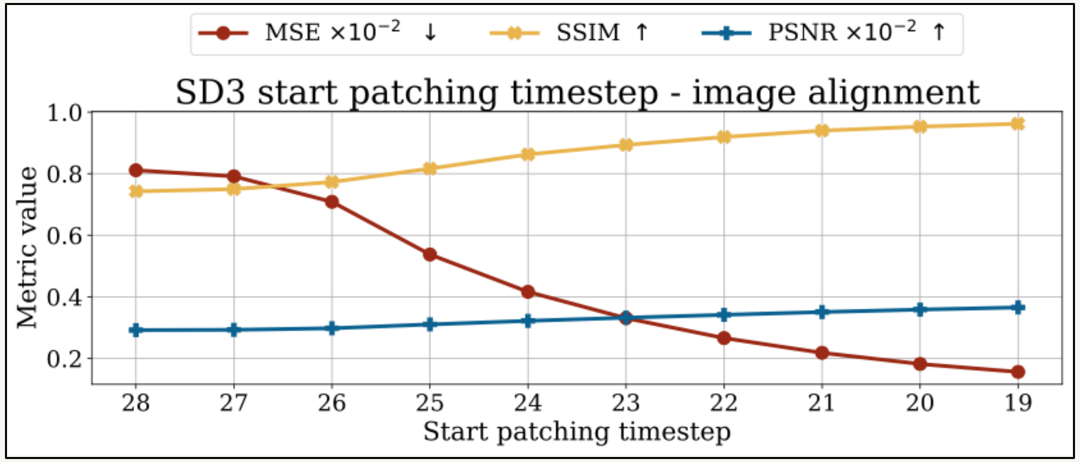
[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

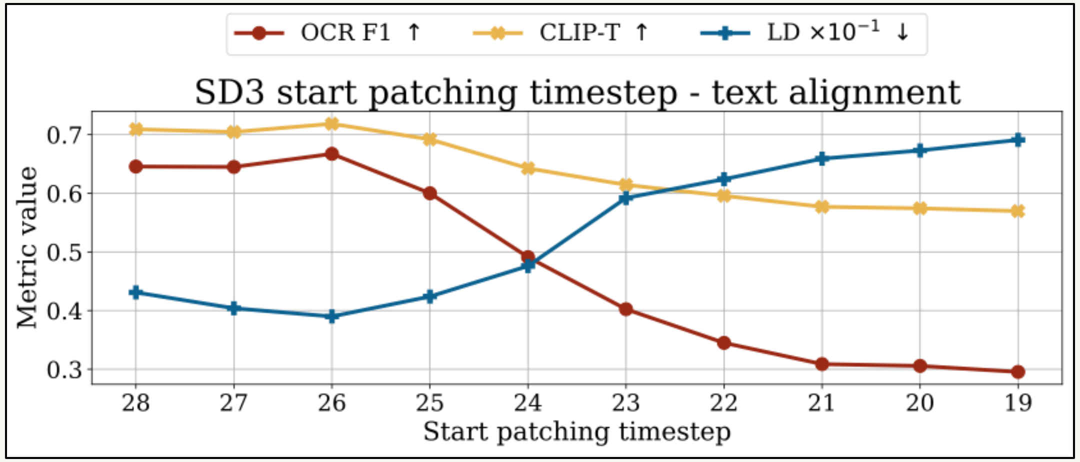
Fine-tuning random layers



[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Does timestep matter in localization?





[6] Staniszewski, Łukasz & Cywiński, Bartosz, et al. "Precise Parameter Localization for Textual Generation in Diffusion Models." ICLR 2025.

Knowledge editing in Text-to-Image models

Removing R2D2

Updating the President of US

Edit: Modify trademarked '*Snoopy*'

Edit: Update with correct '*British Monarch*'