Ready, aim, edit! 🎯 Precise Parameter Localization for Text Editing with Diffusion Models

Łukasz Staniszewski[*1] Bartosz Cywiński[*1]
Franziska Boenisch[2] Kamil Deja[1,3] Adam Dziedzic[2]

[1]Warsaw University of Technology [2]CISPA Helmholtz Center for Information Security [3]IDEAS NCBR

[*]Equal Contribution

#TLDR

- Using attention patching we **localize** a small subset of diffusion models' parameters that determine textual content generated in images.
- We utilize localized parameters in (1) a new method for **image-to-image text edition**, (2) a new text-objective **fine-tuning strategy** and (3) **prevention of toxic text generation**.

Parameter localization

<1%

of DM parameters determine textual content

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%

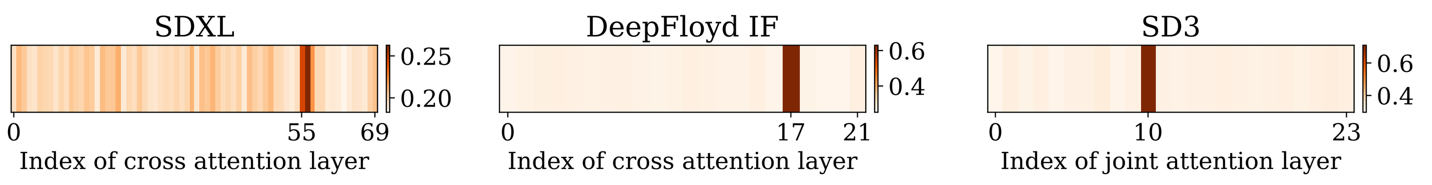


Figure 1. OCR F1-Score with patching of specific attention layers.

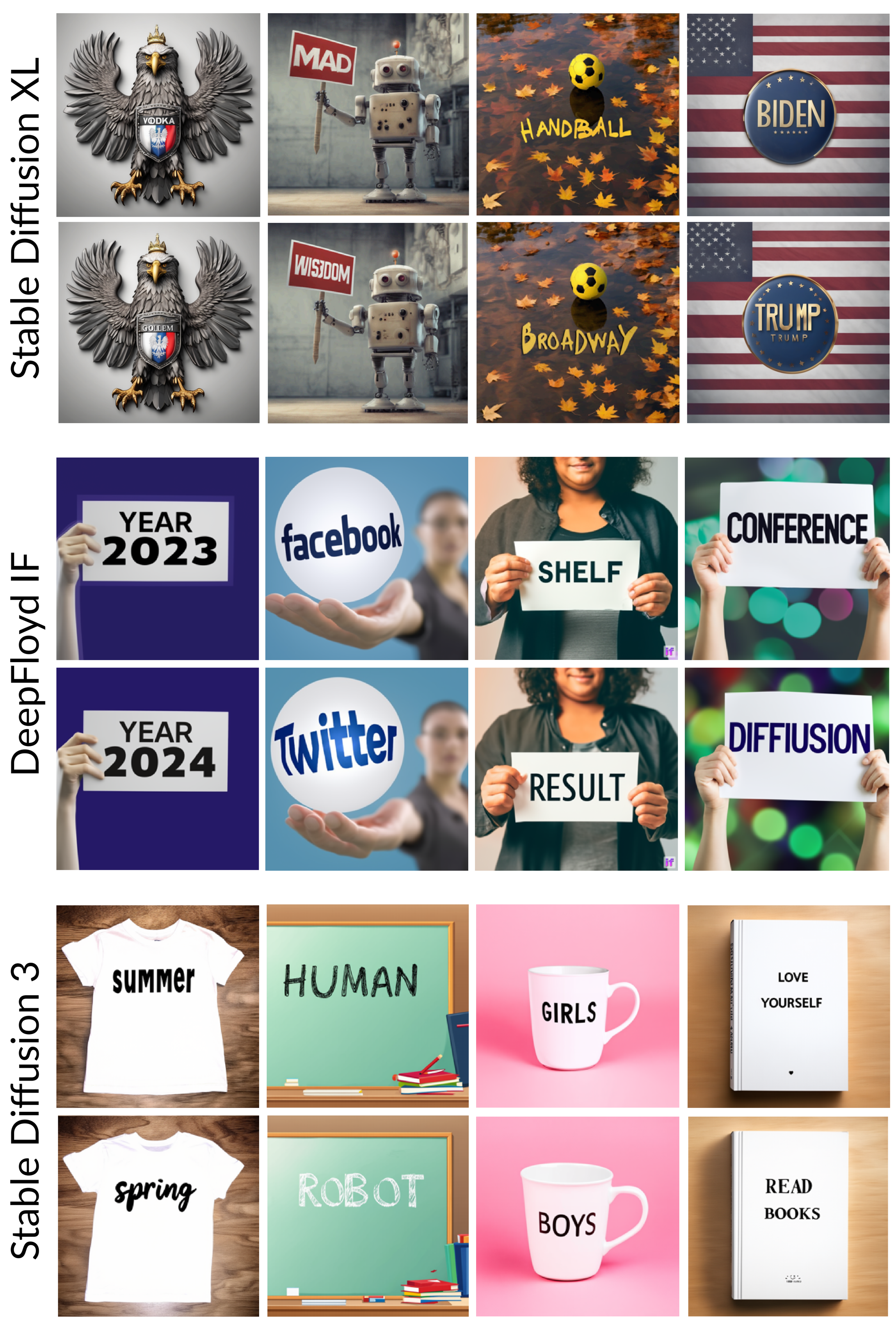
Text-background separation

Localized layers steer the output text while they do not change other visual aspects.

Target Template	Target Text	CLIP-T Template$_S$	CLIP-T Template$_T$	OCR F1 Text$_S$	OCR F1 Text$_T$
Source	Source	0.71	0.44	**0.48**	0.23
Source	Target	0.69	0.44	0.24	0.38
Target	Target	0.71	0.45	0.25	0.36

Table 1. We pass different combinations of templates and keywords texts as a target prompt P_T and show the change in text alignment without major background modifications.

Text editing with our method



Patching vs Injection

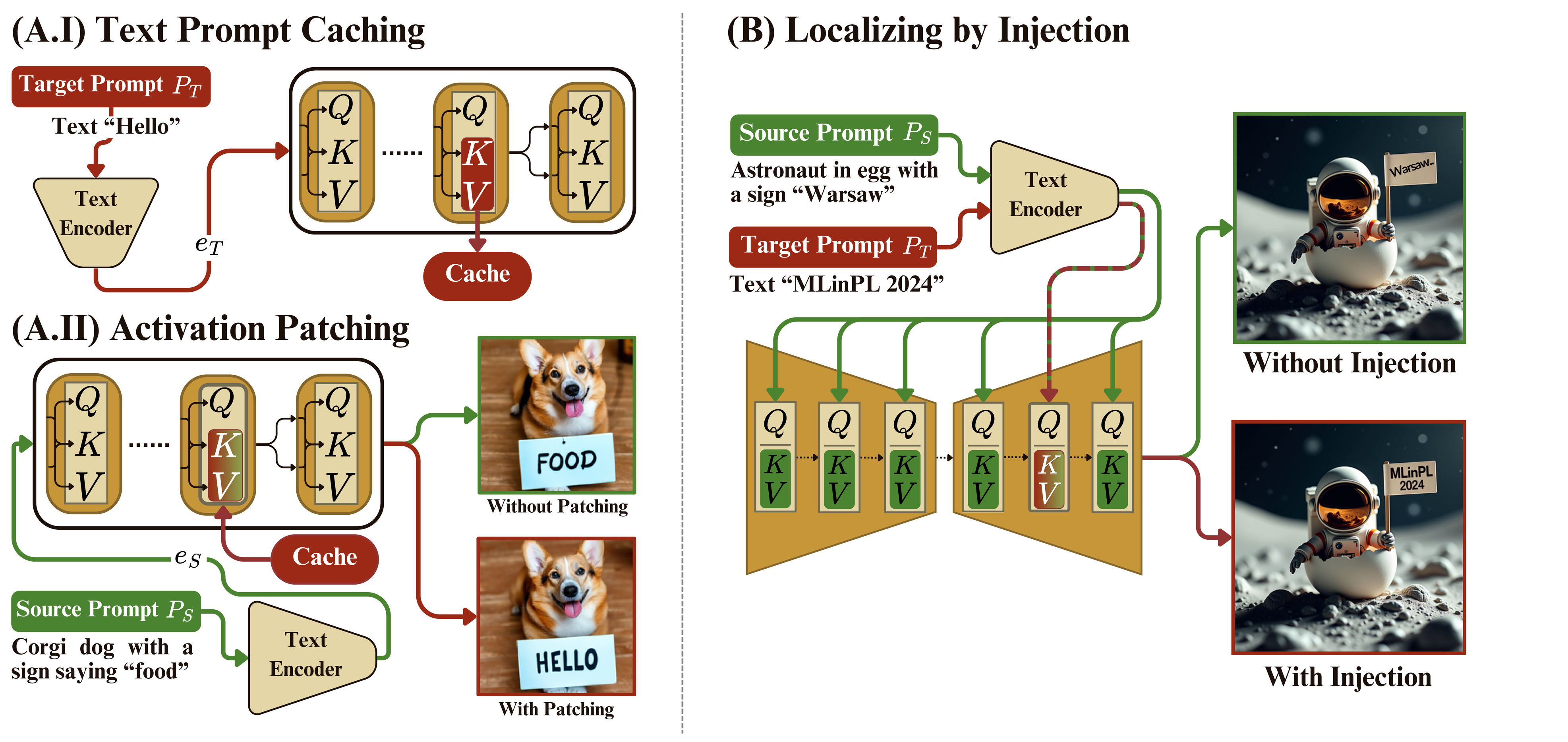


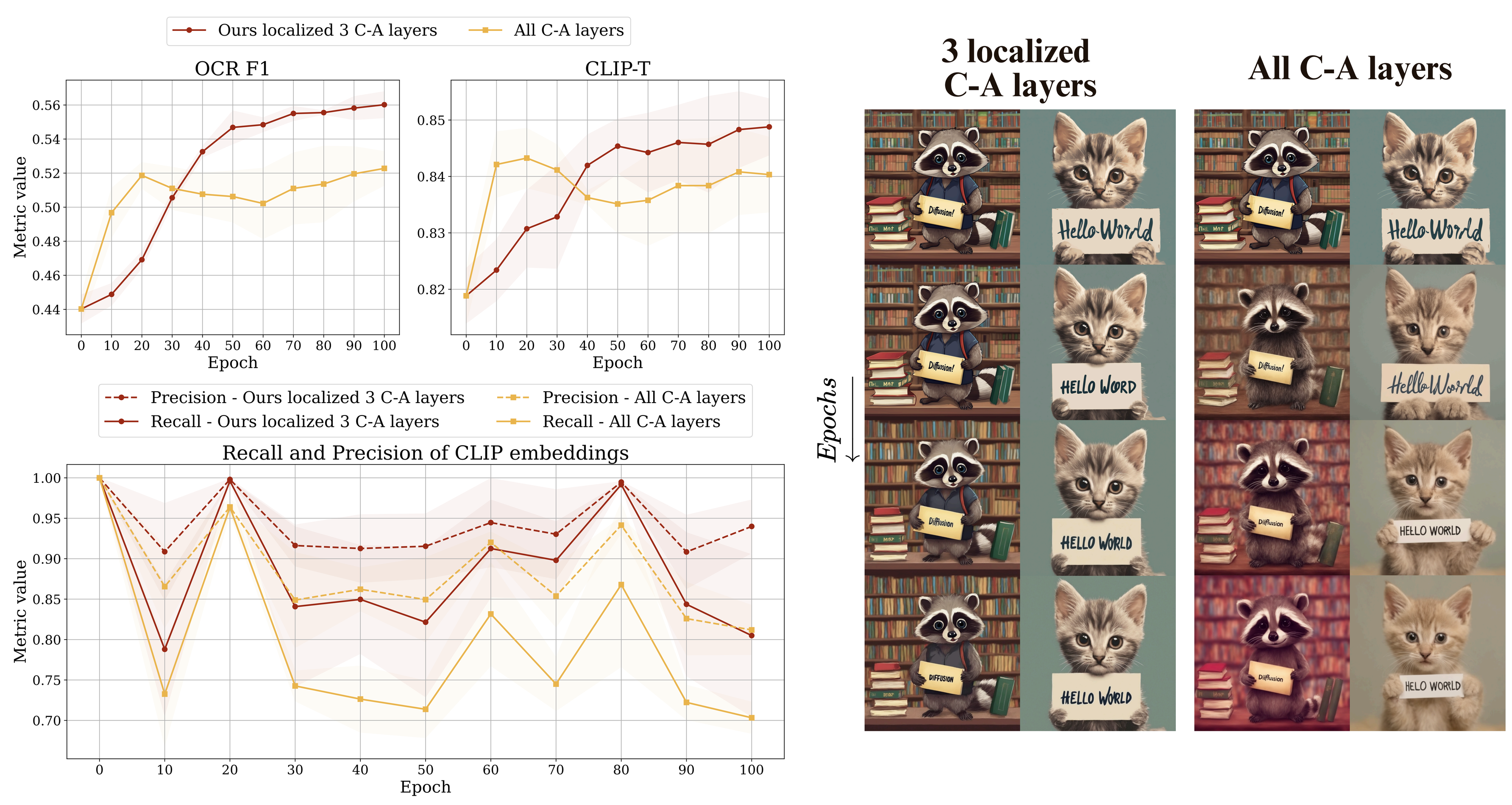
Image-to-image text edition

1. We introduce a new **image-to-image text edition method**, outperforming Prompt-to-Prompt in both background preservation and speed.

Setup	Model	SimpleBench Image alignment SSIM ↑	SimpleBench Image alignment PSNR ↑	SimpleBench Text alignment OCR F1 ↑	SimpleBench Text alignment CLIP-T ↑	CreativeBench Image alignment SSIM ↑	CreativeBench Image alignment PSNR ↑	CreativeBench Text alignment OCR F1 ↑	CreativeBench Text alignment CLIP-T ↑	Execution Time [s] ↓
Ours	SDXL	0.81	32.25	0.34	**0.78**	0.90	35.42	0.32	**0.82**	**10.37**$_{\pm.25}$
Ours LoRA	SDXL	**0.90**	**36.38**	**0.43**	0.77	**0.91**	**37.47**	**0.33**	0.77	**10.37**$_{\pm.25}$
P2P	SDXL	0.82	30.77	0.29	0.69	0.83	30.93	0.26	0.78	31.17$_{\pm.19}$
Ours	IF	**0.64**	**29.90**	**0.70**	**0.81**	**0.74**	31.46	**0.48**	**0.84**	**13.87**$_{\pm.04}$
P2P	IF	0.41	27.90	0.27	0.61	**0.74**	96.84	0.08	0.61	28.04$_{\pm.28}$
P2P*	IF	0.21	27.91	0.41	0.67	0.67	**96.85**	0.11	0.62	28.04$_{\pm.28}$
Ours	SD3	0.72	**29.84**	0.53	0.70	0.73	**30.61**	0.41	0.75	**15.23**$_{\pm.19}$
P2P	SD3	**0.82**	28.65	0.31	0.57	**0.82**	29.13	0.29	0.71	118.30$_{\pm.55}$
P2P*	SD3	0.58	28.24	**0.90**	**0.88**	0.64	28.90	**0.66**	**0.90**	118.30$_{\pm.55}$

Fine-tuning

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



Toxic text prevention

3. Our method successfully **prevents the generation of 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
Prompt Swap	SDXL	0.66	**0.19**	**0.000**
Ours	SDXL	<u>0.79</u>	<u>0.20</u>	<u>0.003</u>
Negative prompt	IF	0.37	0.59	0.250
Safe Diffusion*	IF	**0.74**	0.79	0.540
Prompt Swap	IF	0.35	**0.30**	**0.015**
Ours	IF	<u>0.61</u>	<u>0.32</u>	<u>0.018</u>
Negative prompt	SD3	0.53	0.77	0.407
Safe Diffusion*	SD3	**0.87**	0.73	0.568
Prompt Swap	SD3	0.51	**0.30**	**0.015**
Ours	SD3	<u>0.70</u>	<u>0.32</u>	<u>0.018</u>

Reach out to the authors!

Łukasz Staniszewski Bartosz Cywiński

lukaszstaniszewski10@gmail.com

Supported by PL-Grid Infrastructure grant no. PLG/2024/017266.

bcywinski11@gmail.com