

Skoltech



C>ONSTRUCTOR
UNIVERSITY

Guide-and-Rescale

Self-guidance Mechanism for Effective Tuning-Free Real Image Editing

Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitriy Vetrov, Aibek Alanov

European Conference on Computer Vision, October, 2024



EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO
2024

Paper summary

Introduction, Problem Statement

- Challenge
 - Manipulating real images with t2im models while preserving original structure

Original Image



A photo of a tiger



A photo of a lion

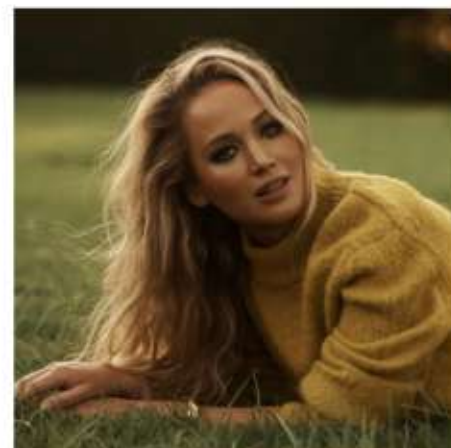


A photo of a goat

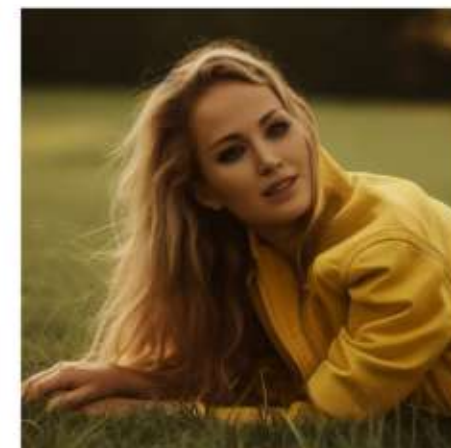


A photo of a wolf

Original Image



Lying down woman wearing yellow sweater



Lying down woman wearing yellow jacket



Lying down woman wearing red sweater



Lying down woman wearing green jacket

Original Image



A photo of a woman



A photo of a sad woman



A photo of a happy woman



A photo of a crying woman

Original Image



A photo



Anime style face



A Pixar style face



A mosaic depicting of a face

Paper summary

Introduction, Baselines

- Existing approaches
 - Fine-tuning (Memory Intensive + Time Consuming)
 - Internal representation exchange (Limited versatility)
 - High-quality reconstruction (Time consuming)

Paper summary

Introduction, Key contributions

- Proposed Framework: Guide-and-Rescale
 - Guidance technique to preserve original structure
 - No-fine tuning or exact inversion required
- Key benefits:
 - Controllable structure preservation via special guidance
 - Automative Noise Rescaling technique
 - Best balance between edit quality and original image preservation

Technical Details

Guidance recap

- Classifier guidance & classifier-free guidance

$$\hat{\epsilon}_t = \epsilon_\theta(z_t; t, y) - s\sigma_t \nabla_{z_t} \log p(y|z_t)$$

$$\hat{\epsilon}_t = (1 + s)\epsilon_\theta(z_t; t, y) - s\epsilon_\theta(z_t; t, \emptyset)$$

- Combined guidance approach

$$\hat{\epsilon}_t = (1 + s)\epsilon_\theta(z_t; t, y) - s\epsilon_\theta(z_t; t, \emptyset) + v\sigma_t \nabla_{z_t} g(z_t; t, y)$$

Technical Details

Proposed guidance scheme

$$\hat{\epsilon}_\theta(z_t, t, y) = \text{CFG}(z_t, t, y_{\text{trg}}, 7.5) + v \cdot \nabla_{z_t} g(z_t, z_t^*, t, y_{\text{src}}, \mathcal{I}^*, \bar{\mathcal{I}})$$

- Self-attention guider

$$g(z_t, z_t^*, t, y_{\text{src}}, \{\mathcal{A}_i^{\text{self}}\}, \{\bar{\mathcal{A}}_i^{\text{self}}\}) = \sum_{i=1}^L \text{mean} \|\mathcal{A}_i^{\text{self}} - \bar{\mathcal{A}}_i^{\text{self}}\|_2^2$$

- Feature guider

$$g(z_t, z_t^*, t, y_{\text{src}}, \Phi^*, \bar{\Phi}) = \text{mean} \|\Phi^* - \bar{\Phi}\|_2^2$$

Technical Details

Proposed guidance scheme, motivation

Source prompt: "A photo of a woman wearing a shirt with a drawing"
Target prompt: "A photo of a woman wearing a red shirt with a drawing"

Initial image



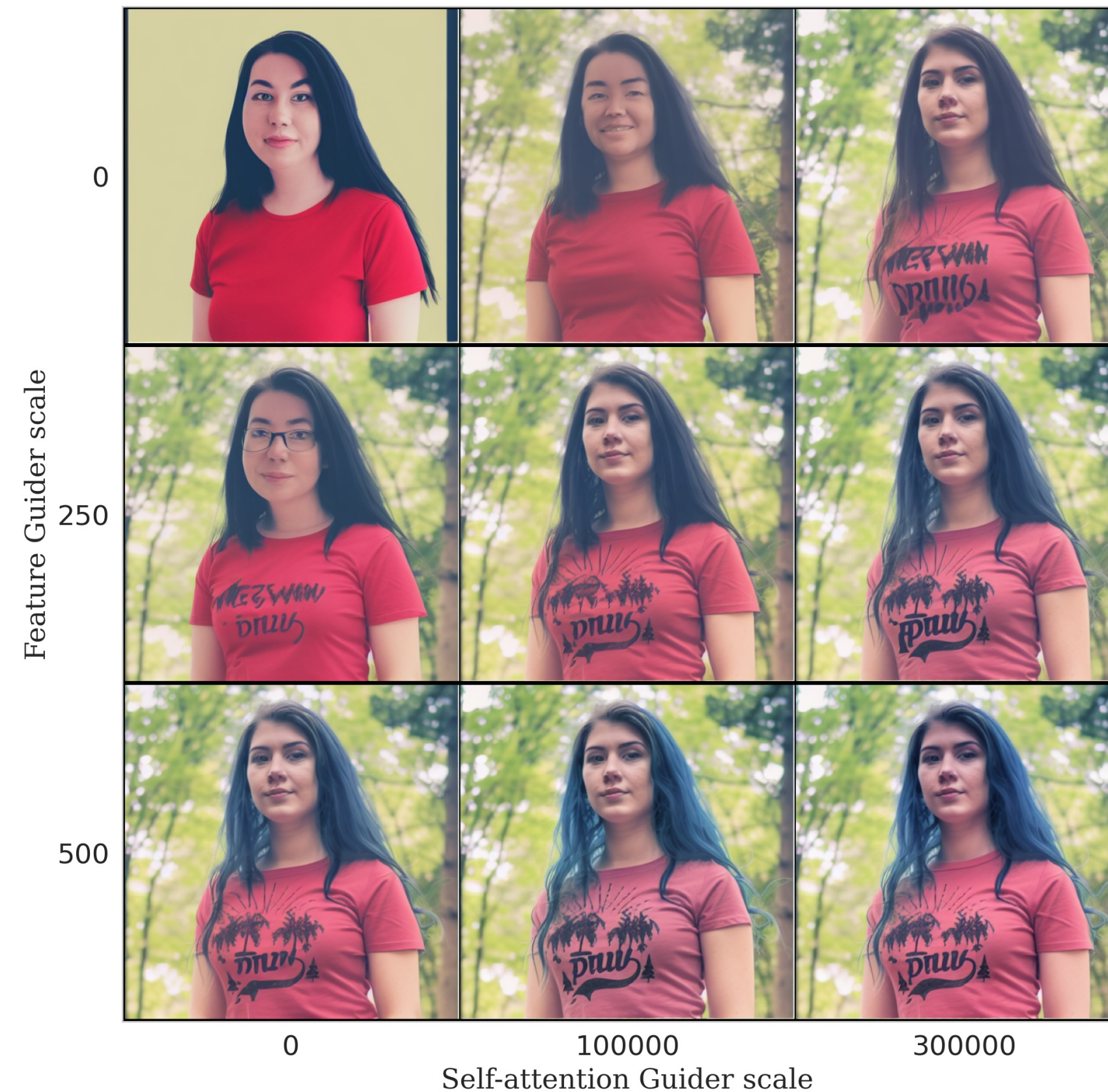
Naive editing



L2 Guider



Our editing



Technical Details

Rescaling technique

- Overall guidance scheme

$$\epsilon_t = \epsilon_\theta(z_t, t, \emptyset) + w(\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \emptyset)) + \gamma \sum_i v_i \cdot \nabla_{z_t} g_i(z_t, z_t^*, t, y_{\text{src}}, \mathcal{I}^*, \bar{\mathcal{I}}),$$

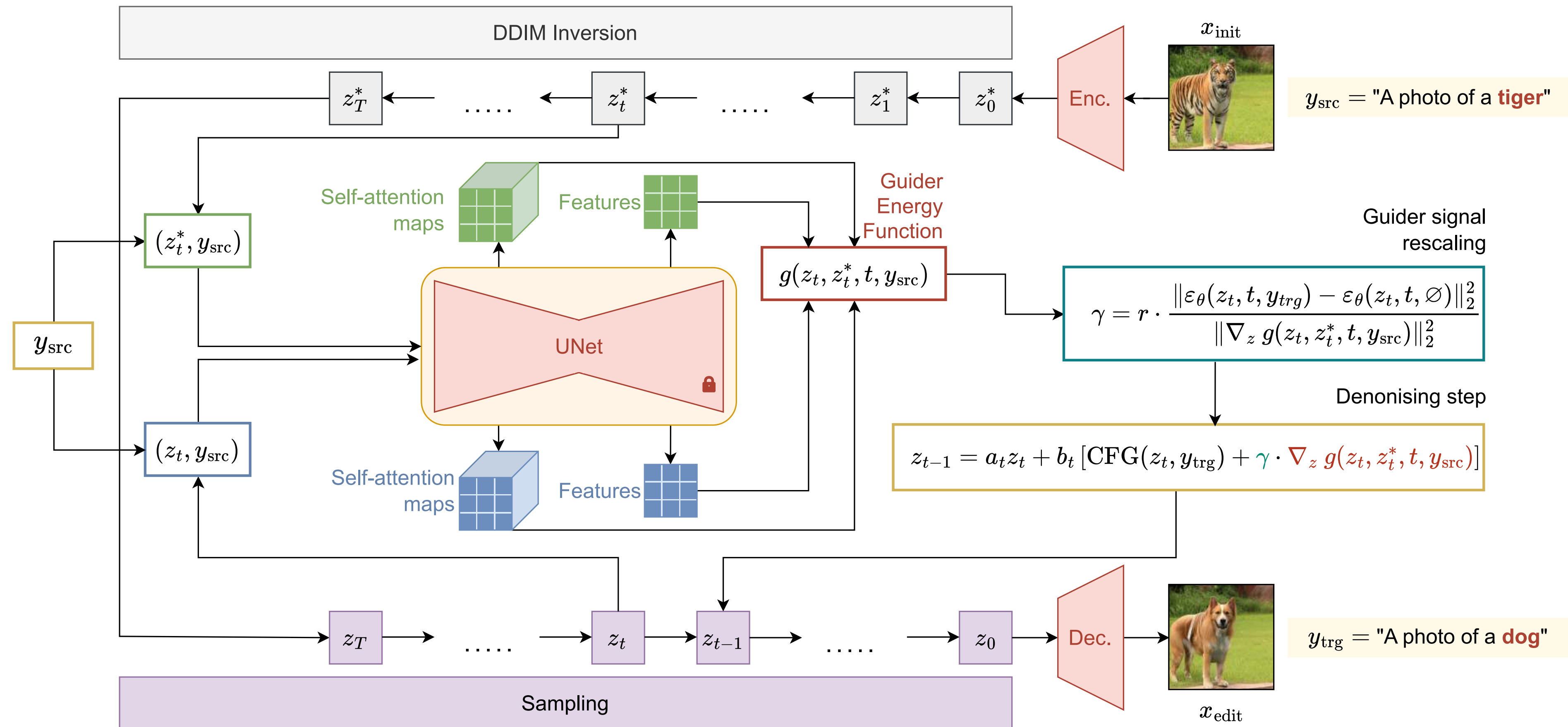
- Automative rescaling scheduling

$$r_{\text{cur}}(t) = \frac{\|w(\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \emptyset))\|_2^2}{\|\sum_i v_i \cdot \nabla_{z_t} g_i(z_t, z_t^*, t, y_{\text{src}})\|_2^2},$$

$$\gamma = r \cdot r_{\text{cur}}(t).$$

Technical Details

Overall pipeline



Results

Qualitative



Results

Quantitative

Method	LPIPS ↓	CLIP ↑	FID ↓	Time (s) ↓
ProxMasaCtrl [5]	0.267	0.215	94.53	12.94
MasaCtrl [2]	0.306	0.223	100.62	13.73
EDICT [27]	<u>0.221</u>	0.229	47.13	68.13
P2P [6] + NTI [15]	0.279	0.233	42.46	66.77
P2P [6] + NPI [14] Prox [5]	0.170	0.233	43.16	8.59
P2P [6] + NPI [14]	0.251	0.234	44.05	8.54
PnP [24]	0.366	0.256	<u>39.55</u>	197.0
Guide-and-Rescale (ours)	0.228	<u>0.243</u>	39.07	24.26

Contacts

HF Demo:



Code:



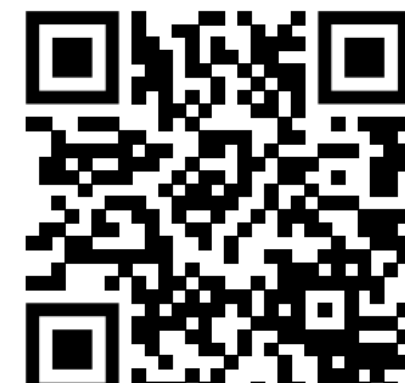
Project page:



Vadim
Titov



Madina
Khalmatova



Alexandra
Ivanova



Dmitry
Vetrov



Aibek
Alanov

