MagicTailor: Component-Controllable Personalization in Text-to-Image Diffusion Models

Donghao Zhou[1*], Jiancheng Huang[2*], Jinbin Bai[3], Jiaze Wang[1], Hao Chen[1], Guangyong Chen[4], Xiaowei Hu[5†], Pheng-Ann Heng[1]
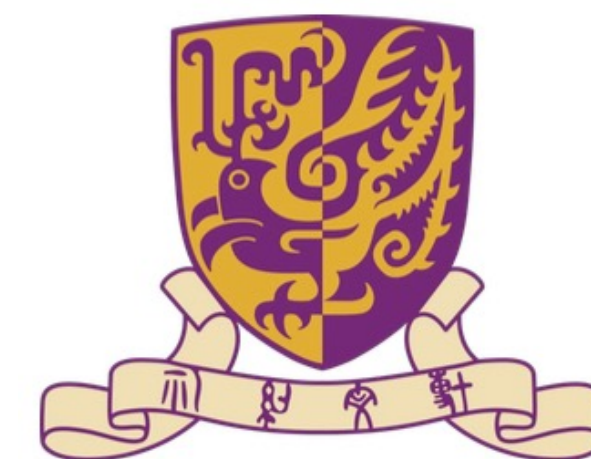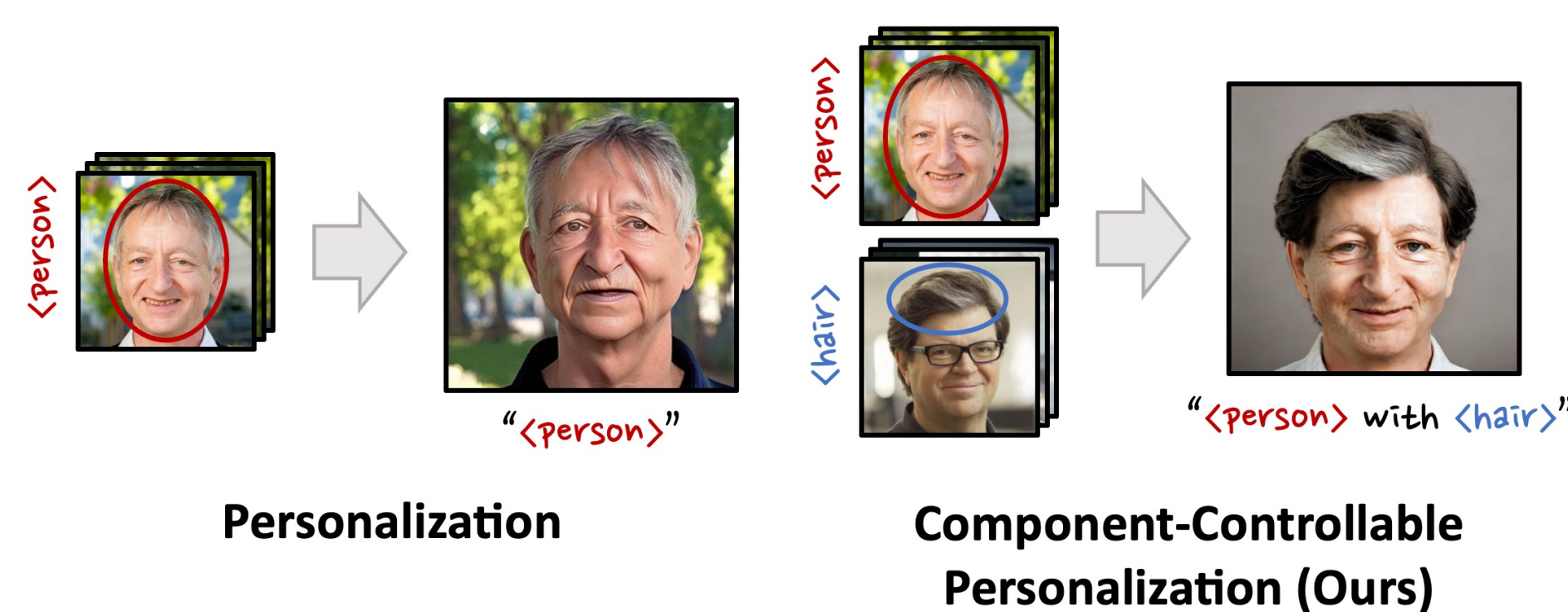
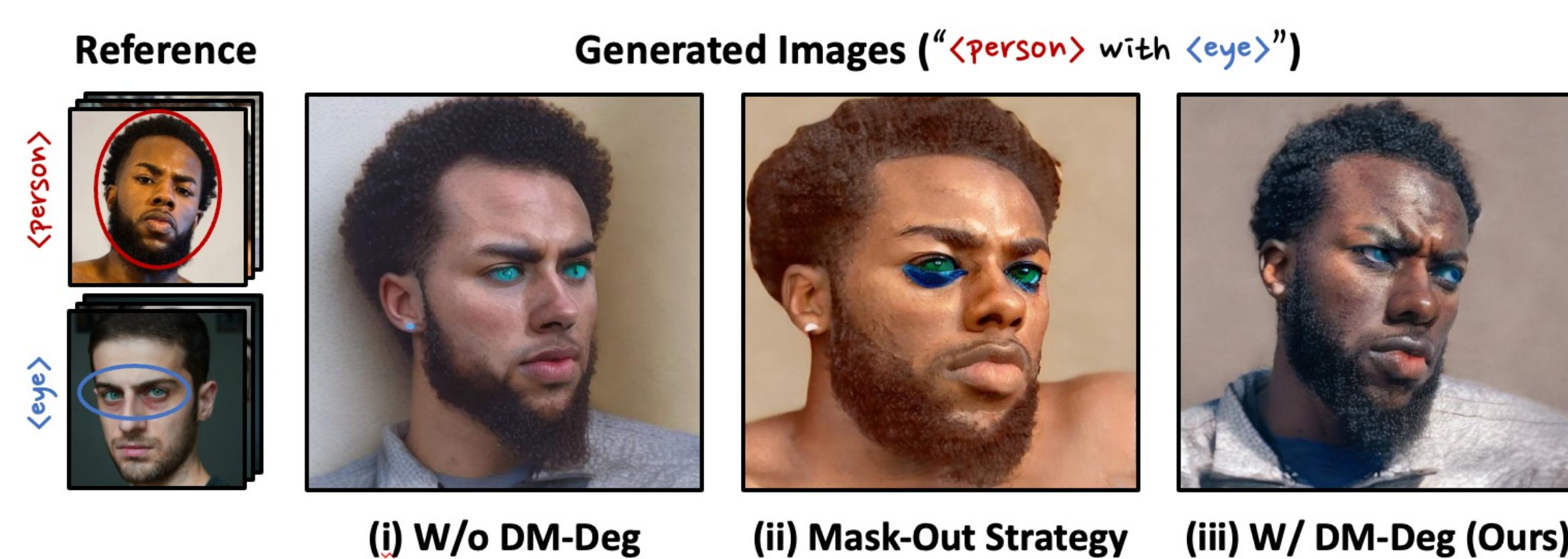[1]CUHK [2]SIAT, CAS, [3]NUS, [4]Zhejiang Lab, [5]Shanghai AI Lab

Task Formulation

We study **component-controllable Personalization**, a new task aiming to reconfigure specific components of concepts during personalization.
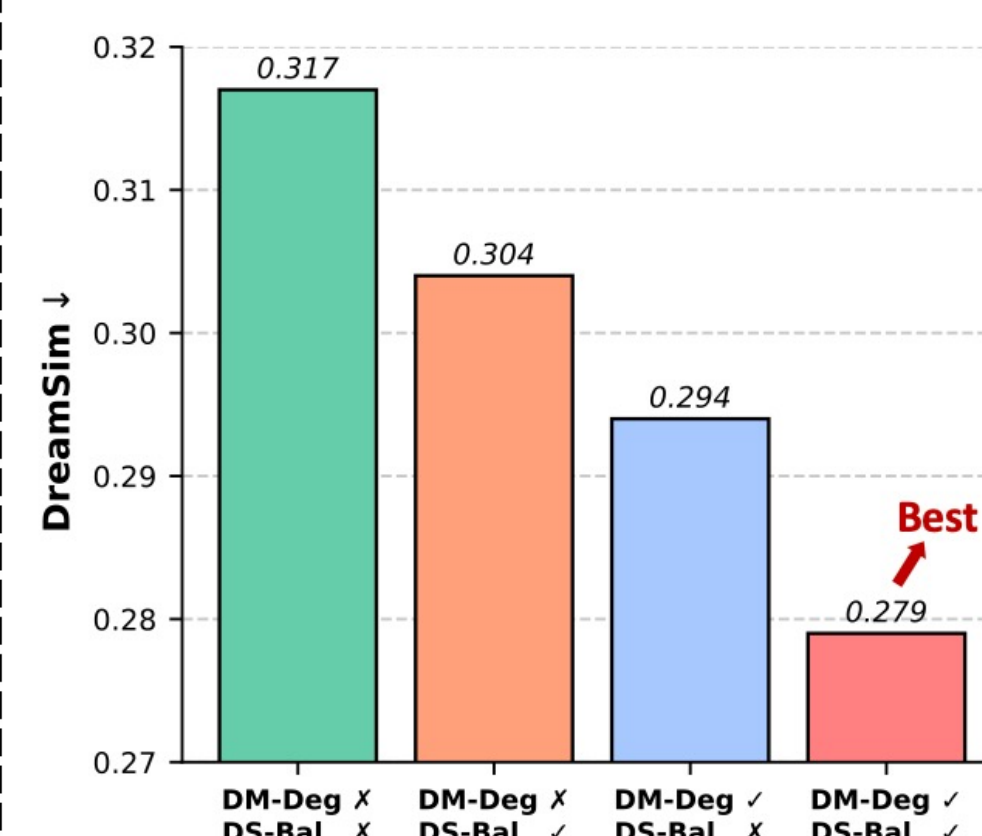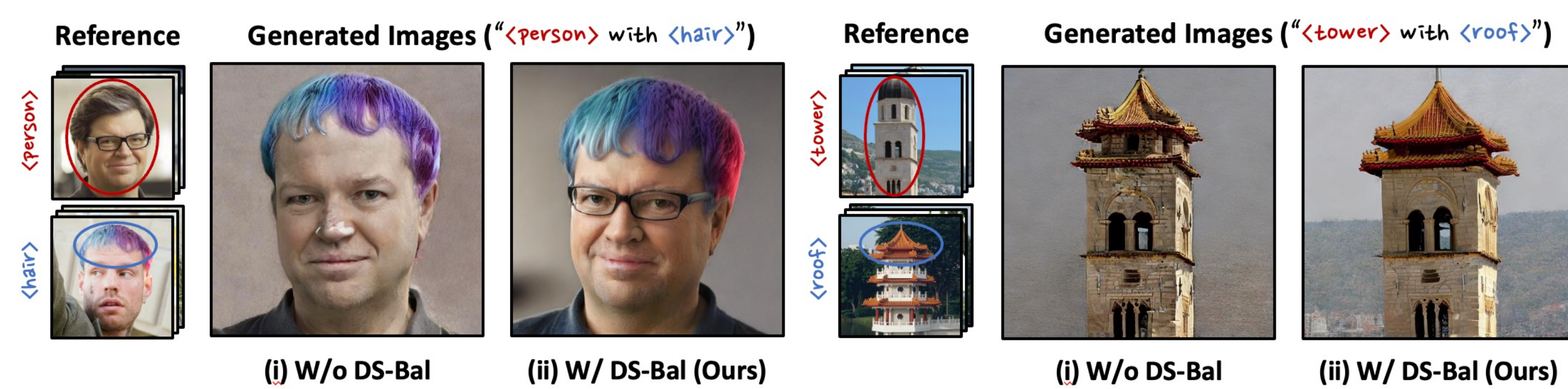

Personalization

Component-Controllable Personalization (Ours)

Challenges

(a) Challenge #1: Semantic Pollution

Reference | Generated Images ("⟨person⟩ with ⟨eye⟩")

(i) W/o DM-Deg | (ii) Mask-Out Strategy | (iii) W/ DM-Deg (Ours)

(C) Id. Fidelity Performance

DreamSim values: 0.317, 0.304, 0.294, 0.279 (Best)
DM-Deg / DS-Bal combinations

(b) Challenge #2: Semantic Imbalance

Reference | Generated Images ("⟨person⟩ with ⟨hair⟩")
(i) W/o DS-Bal | (ii) W/ DS-Bal (Ours)

Reference | Generated Images ("⟨tower⟩ with ⟨roof⟩")
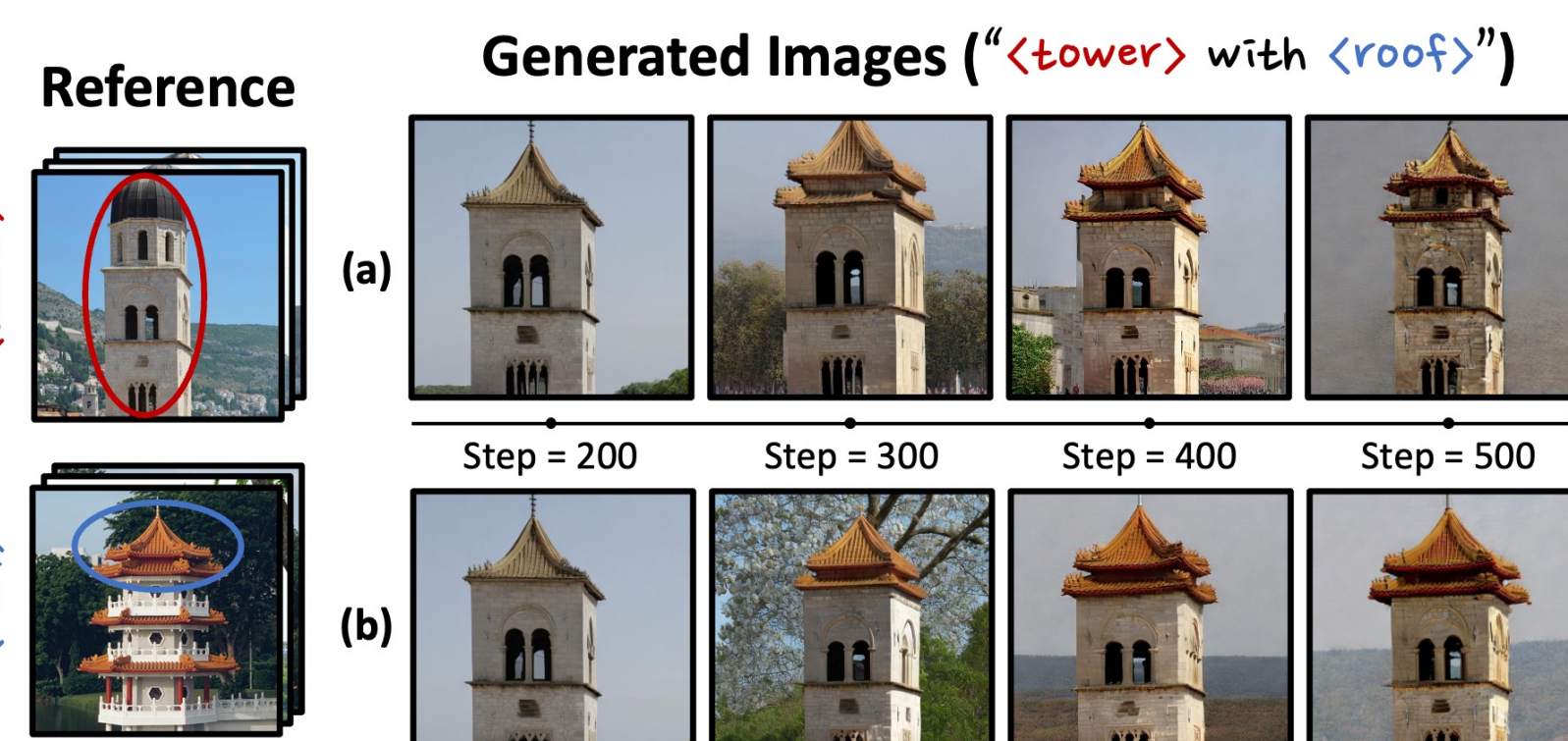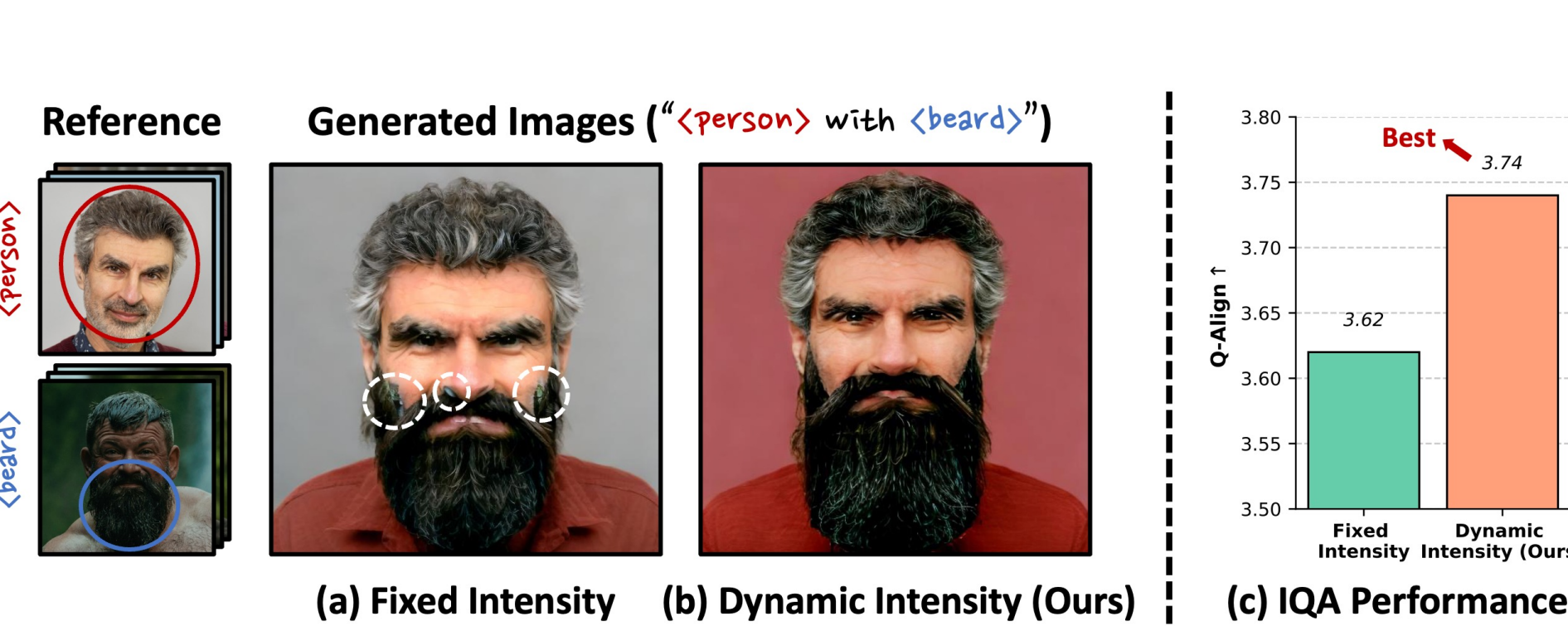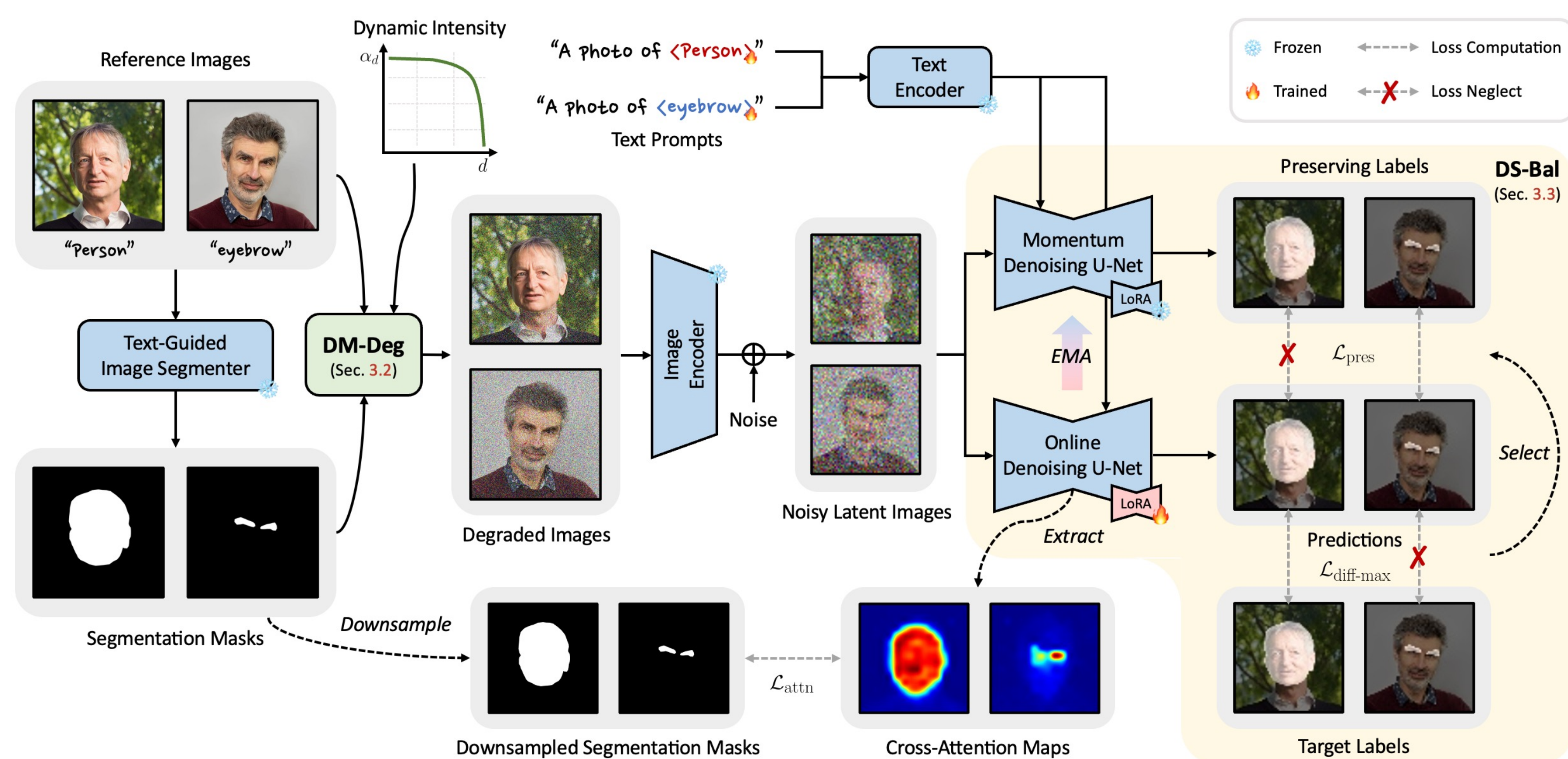(i) W/o DS-Bal | (ii) W/ DS-Bal (Ours)

Semantic Pollution: Unwanted visual elements inadvertently appear in generated images, "polluting" the personalized concept!
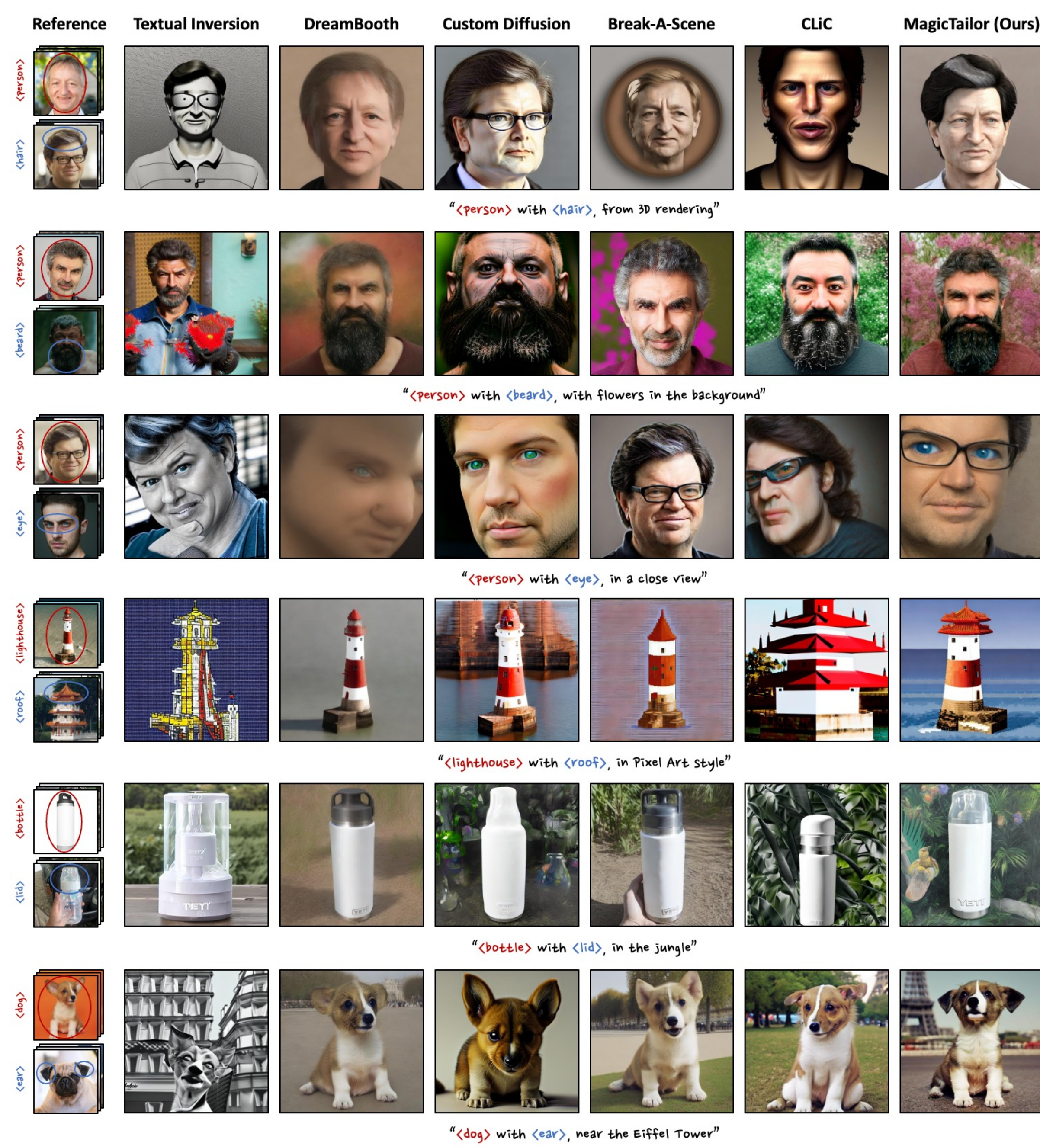
Semantic Imbalance: The model could overemphasizes the concept or the component, leading to unfaithful personalization!

Methodology

We propose **MagicTailor**, an innovative framework consisting of **Dynamic Masked Degradation (DM-Deg)** and **Dual-Stream Balancing (DS-Bal)**.

(a) Fixed Intensity | (b) Dynamic Intensity (Ours) | (c) IQA Performance

Q-Align: 3.62 (Fixed Intensity), 3.74 (Dynamic Intensity) Best

Reference | Generated Images ("⟨tower⟩ with ⟨roof⟩")
Step = 200 | Step = 300 | Step = 400 | Step = 500

Experiments

Reference | Textual Inversion | DreamBooth | Custom Diffusion | Break-A-Scene | CLiC | MagicTailor (Ours)

"⟨person⟩ with ⟨hair⟩, from 3D rendering"
"⟨person⟩ with ⟨beard⟩, with flowers in the background"
"⟨person⟩ with ⟨eye⟩, in a close view"
"⟨lighthouse⟩ with ⟨roof⟩, in Pixel Art style"
"⟨bottle⟩ with ⟨lid⟩, in the jungle"
"⟨dog⟩ with ⟨ear⟩, near the Eiffel Tower"

Qualitative Comparison

Methods	CLIP-T ↑	CLIP-I ↑	DINO ↑	DreamSim ↓
Textual Inversion [Gal *et al.*, 2022]	0.236	0.742	0.620	0.558
DreamBooth [Ruiz *et al.*, 2023]	0.266	0.841	0.798	0.323
Custom Diffusion [Kumari *et al.*, 2023]	0.251	0.797	0.750	0.407
Break-A-Scene [Avrahami *et al.*, 2023]	0.259	0.840	0.780	0.338
CLiC [Safaee *et al.*, 2024]	0.263	0.764	0.663	0.499
MagicTailor (Ours)	**0.270**	**0.854**	**0.813**	**0.279**

Quantitative Comparison

DM-Deg	DS-Bal	CLIP-T ↑	CLIP-I ↑	DINO ↑	DreamSim ↓
		0.275	0.837	0.798	0.317
✓		**0.276**	0.848	0.809	0.294
	✓	0.270	0.845	0.802	0.304
✓	✓	0.270	**0.854**	**0.813**	**0.279**

Ablation Studies

Reference | Generated Images
"⟨person⟩ with ⟨hair⟩" | "⟨person⟩ with ⟨ear⟩"

Generalizability for Difficult Pairs

"⟨dog⟩ with ⟨ear⟩, joyful, running on a sunny day"
"⟨dog⟩ with ⟨ear⟩, calm, lying down, on a wooden floor"
"⟨dog⟩ with ⟨ear⟩, holding a red ball in the mouth, on the bench"
"⟨person⟩ with ⟨beard⟩, focused, playing guitar, in an outdoor field"
"⟨person⟩ with ⟨beard⟩, grabbing a cup, smiling, in a coffee shop"
"⟨person⟩ with ⟨beard⟩, looking serious, wearing glasses, in a library"
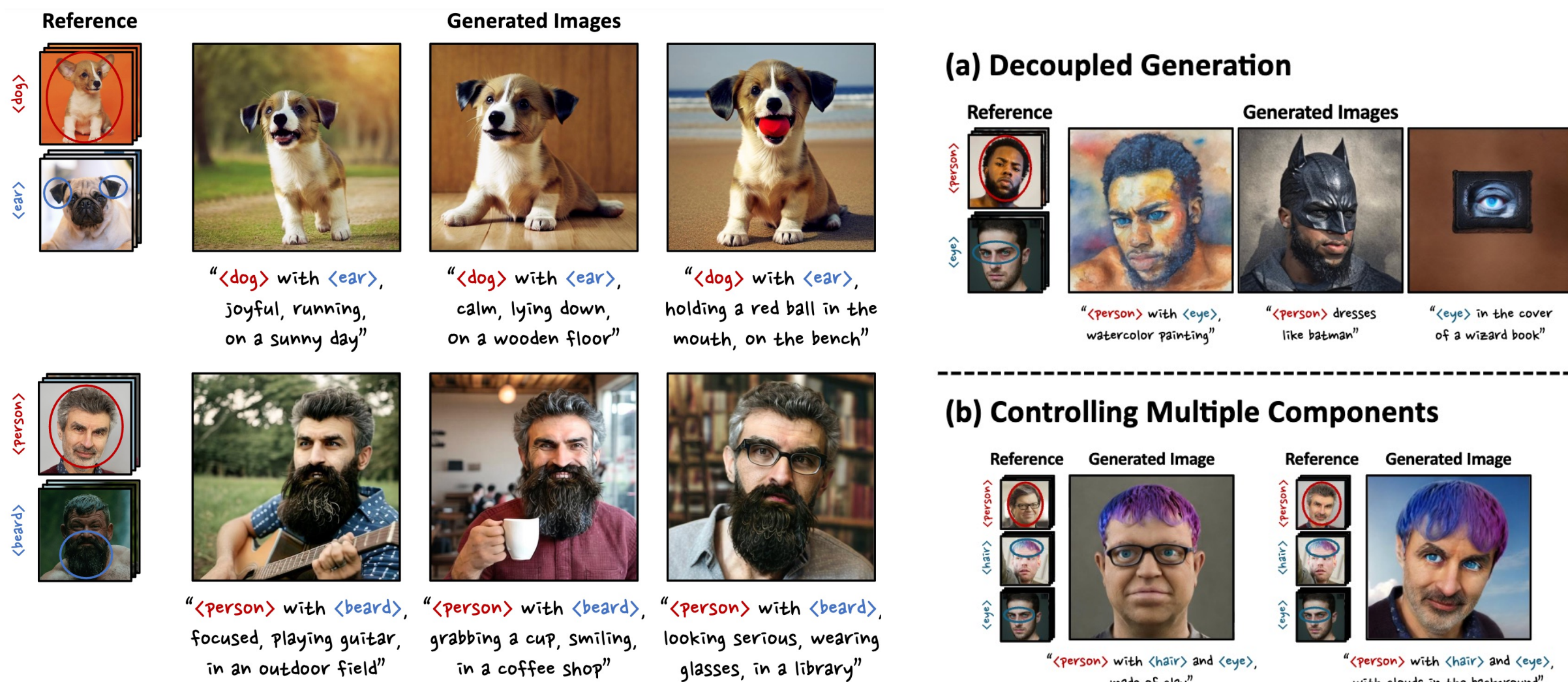
Generalizability for Complex Prompts

(a) Decoupled Generation
(b) Controlling Multiple Components

Further Applications I

(c) Enhancing Other Generative Tools

MagicTailor + ControlNet
MagicTailor + CSGO
MagicTailor + InstantMesh

Further Applications II