# EZIGen: Enhancing zero-shot personalized image generation with precise subject encoding and decoupled guidance

Zicheng Duan[1], Yuxuan Ding[2], Chenhui Gou[3], Ziqin Zhou[1], Ethan Smith[4], Lingqiao Liu[1]

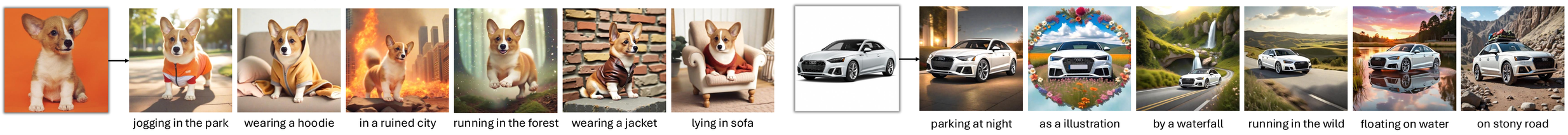
[1]The University of Adelaide, Australia [2]Qualcomm AI Research, Beijing [3]Monash University, Australia [4]Leonardo.AI

BMVC 2025

## Background

**Subject-driven personalized image generation**

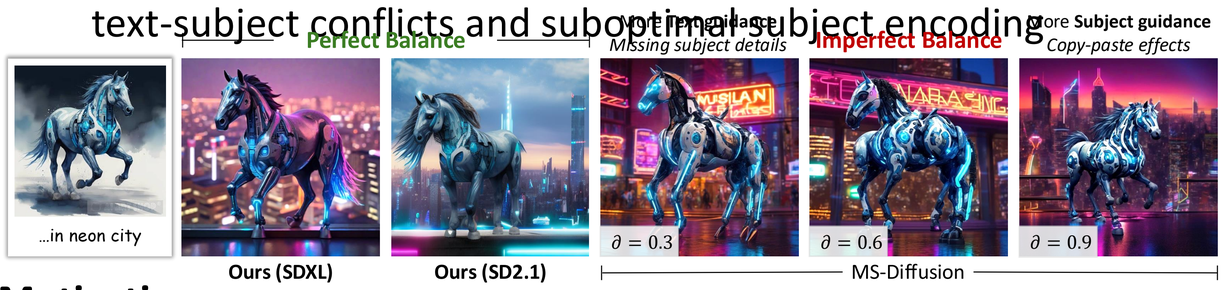
- Subject appearance from given subject image
- Overall semantic from input text prompt

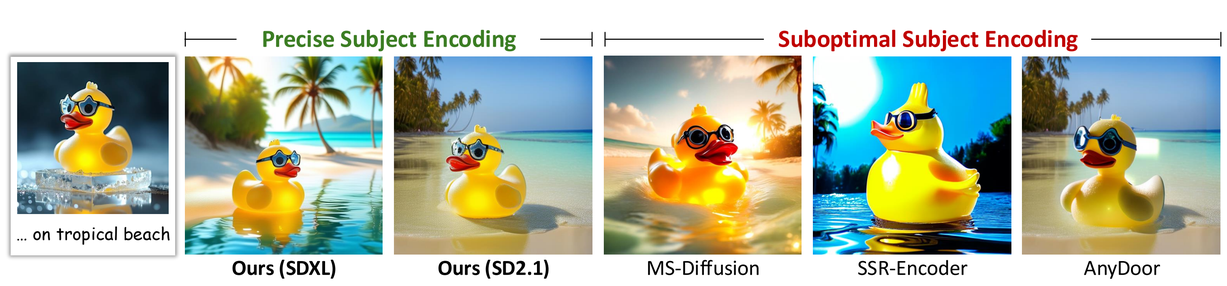



## Motivation

**Related works:**

1) Text to image generation: Generates images from given text prompt *However*, lack the ability to condition on user-specific subject image
2) Tuning-based personalized image generation: Learn sets of parameters for each subjects. *However*, requires re-training for new subjects.
3) Zero-shot personalized image generation: Generate personalized images condition on arbitrary subject images. *However*, suffering from text-subject conflicts and suboptimal subject encoding

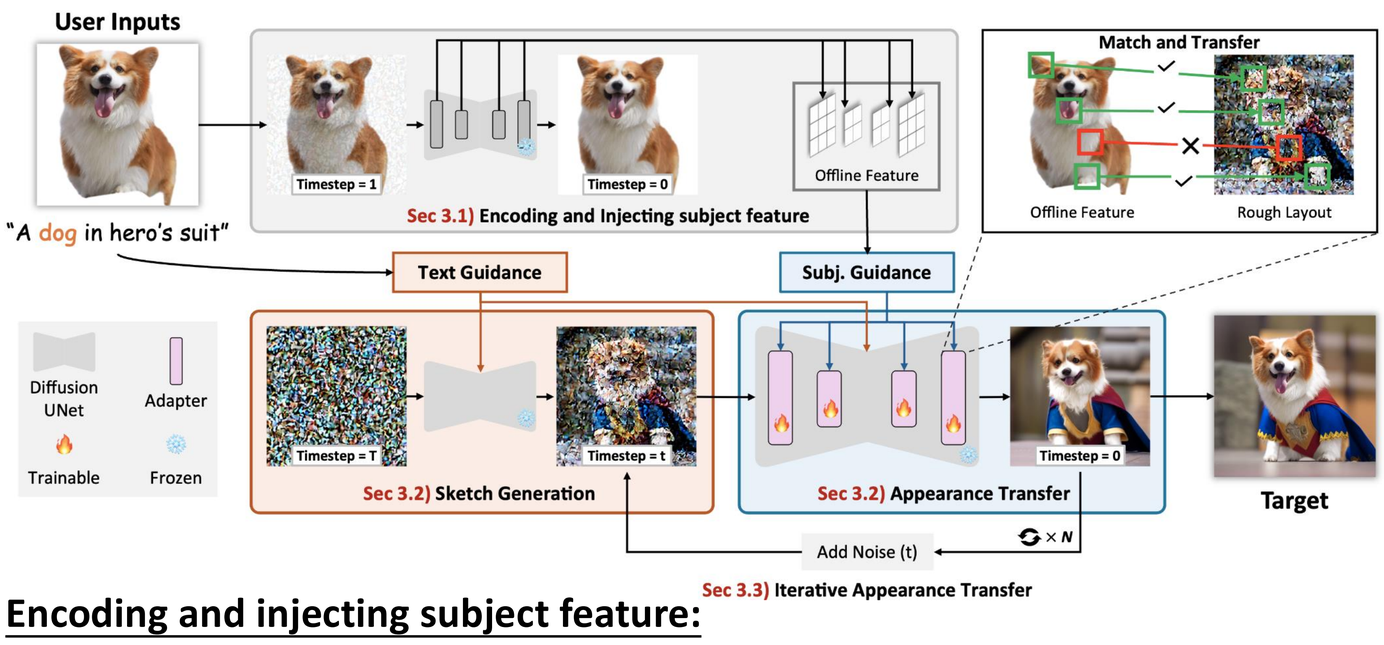



**Motivation:**

Separate textual/subject control and carefully choose a stronger subject encoder, ensuring both subject encoding and textual alignment.

## Method: EZIGen




**Encoding and injecting subject feature:**

**Reference UNet for subject encoding:**

We extract fine-grained subject feature from a fixed pretrained diffusion UNet by adding slight noise and denoise to a subject image $I^{\text{sub}}$ obtain subject feature $\mathcal{F}^{\text{sub}}$:

$$I^{\text{sub}}_{t=1} = I^{\text{sub}} + \phi(\varepsilon, t=1),\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad \mathcal{F}^{\text{sub}} = \{s_1, s_2, \ldots, s_N\} = U(I^{\text{sub}}_{t=1}, t=1)$$

$\{s_1, s_2, \ldots, s_N\}$ represents subject feature extracted from UNet layers.

**Injecting subject feature using adapter:**

$s_n \in \mathcal{F}^{\text{sub}}$ is then injected to diffusion UNet by converting each to key-value pairs with newly registered adapters $A_n$ with concatenated attention:

$$\mathcal{F}_n = A_n(Q_n, [K_n \| K^{\text{sub}}_n], [V_n \| V^{\text{sub}}_n]) \quad K^{\text{sub}}_n = W_k s_n, \quad V^{\text{sub}}_n = W_v s_n$$

$\mathcal{F}_n$ is then passed on to continue generating.

**Decoupling text and subject guidance:**

**Sketch generation process:**

Using pure text prompt to generate a rough intermediate $(T \to t)$ latent containing rough semantic information and general layout, referred as the Sketch of the final image.

**Appearance transfer process:**

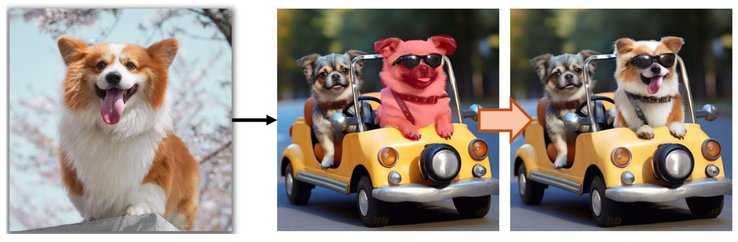
The generation process continues $(t \to 0)$ based on the Sketch and the subject feature $\mathcal{F}^{\text{sub}}$, via the trained adapters $A_N$, aiming to transfer the appearance of subject to the sketch latent for determine the final image $x_0$.

**Iterative appearance transfer**

The overall quality can be further improved by re-adding noise to $x_0$ to obtain a new Sketch and repeat the appearance transfer process, based on that the new Sketch has better visual correspondence with the subject.

**Zero-shot personalized image editing**

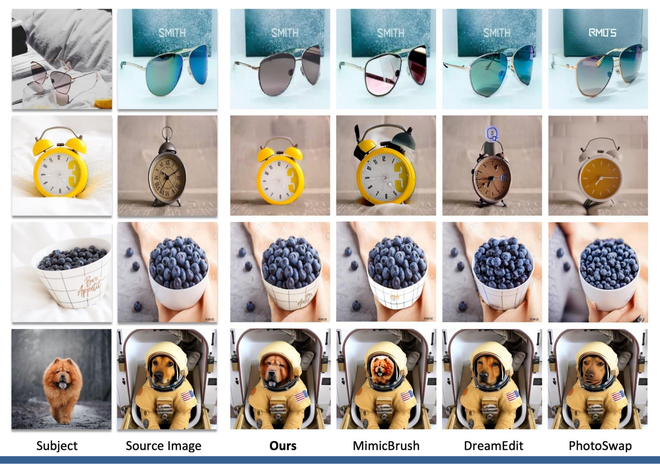
Our method achieves personalized editing when incorporating with image mask and inversion.



## Experiments

**Comparison on personalized image generation**


**Comparison on personalized editing**




**More Visualization Results**