# CCM: Real-Time Controllable Visual Content Creation Using Text-to-Image Consistency Models

Jie Xiao[1], Kai Zhu[2], Han Zhang[3], Zhiheng Liu[1], Yujun Shen[4], Zhantao Yang[3], Ruili Feng[2], Yu Liu[2], Xueyang Fu[1], Zheng-Jun Zha[1]
[1]MPC Lab, University of Science and Technology of China
[2]Alibaba Group [3]Shanghai Jiao Tong University [4]Ant Group

ICML 2024

# Content

- Introduction
- Method
- Experiment
- Conclusion

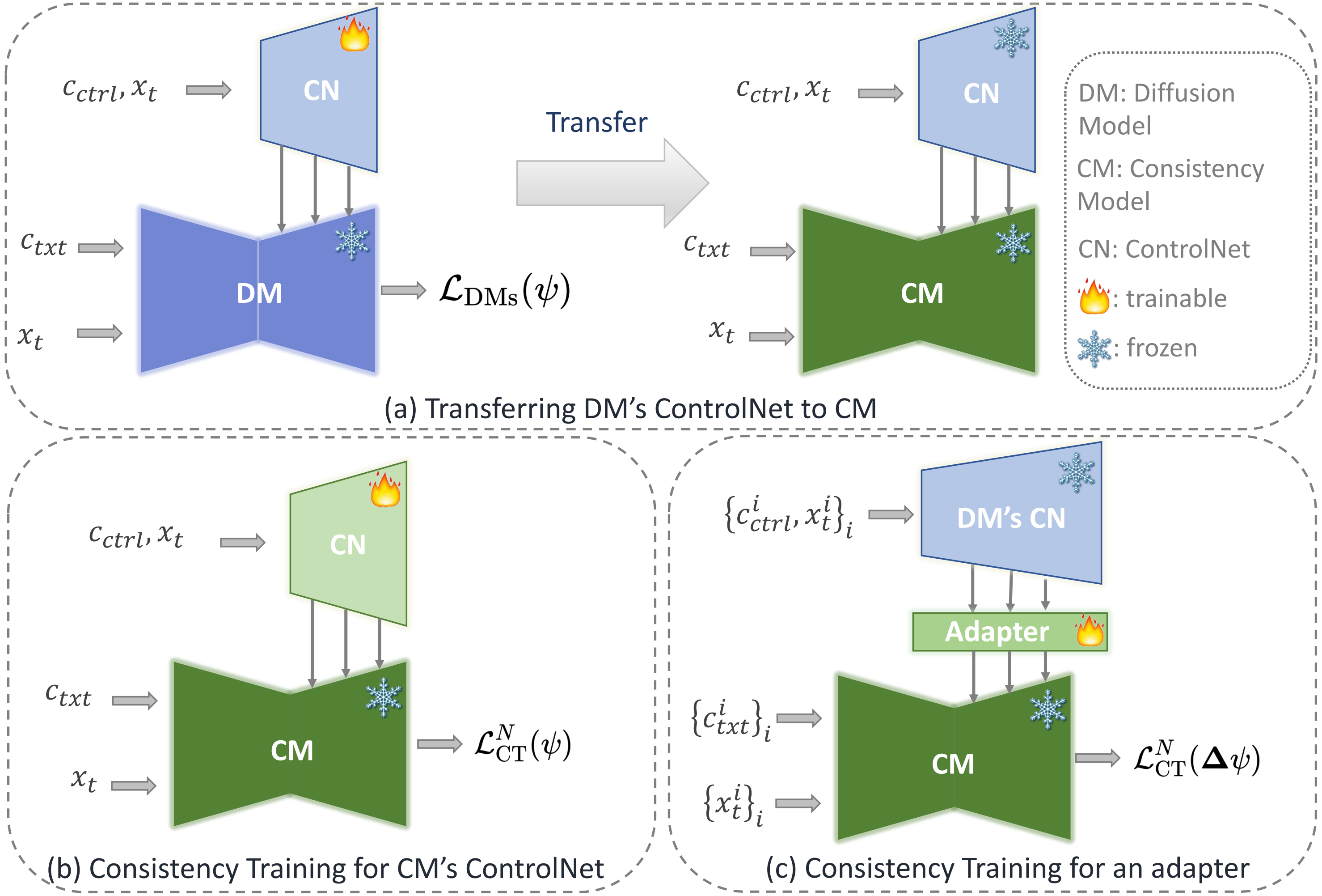
# Introduction

# Motivation

| | | | |
|---|---|---|---|
| Diffusion Model | Quality | Controllable | Efficiency |
| Consistency Model | Quality | Controllable | Efficiency |

# Method

# Overview




(a) Training a ControlNet based on the text-to-image diffusion model (DM) and directly applying it to the text-to-image consistency model (CM); (b) consistency training for ControlNet based on the text-to-image consistency model; (c) consistency training for a unified adapter to utilize better transfer of DM's ControlNet.

# Method

## Applying ControlNet of Text-to-Image Diffusion Models

$$\mathcal{L}_{\mathrm{DMs}}(\psi) = \mathbb{E}\left[\left\|\epsilon - \epsilon_{\{\phi,\psi\}}(x_t, t, c_{\mathrm{txt}}, c_{\mathrm{ctrl}})\right\|_2^2\right]$$

## Consistency Training for ControlNet

$$\mathcal{L}_{\mathrm{CT}}^{N}(\psi) = \mathbb{E}[\lambda(t_n) d(f_{\{\theta,\psi\}}(x_{t_{n+1}}, t_{n+1}; c_{\mathrm{txt}}, c_{\mathrm{ctrl}}), f_{\{\theta,\psi\}^-}(x_{t_n}, t_n; c_{\mathrm{txt}}, c_{\mathrm{ctrl}}))]$$

## Consistency Training for A Unified Adapter

$$\mathcal{L}_{\mathrm{CT}}^{N}(\Delta\psi) = \mathbb{E}[\lambda(t_n) d(f_{\{\theta,\psi,\Delta\psi\}}(x_{t_{n+1}}, t_{n+1}; c_{\mathrm{txt}}, c_{\mathrm{ctrl}}), f_{\{\theta,\psi,\Delta\psi\}^-}(x_{t_n}, t_n; c_{\mathrm{txt}}, c_{\mathrm{ctrl}}))]$$

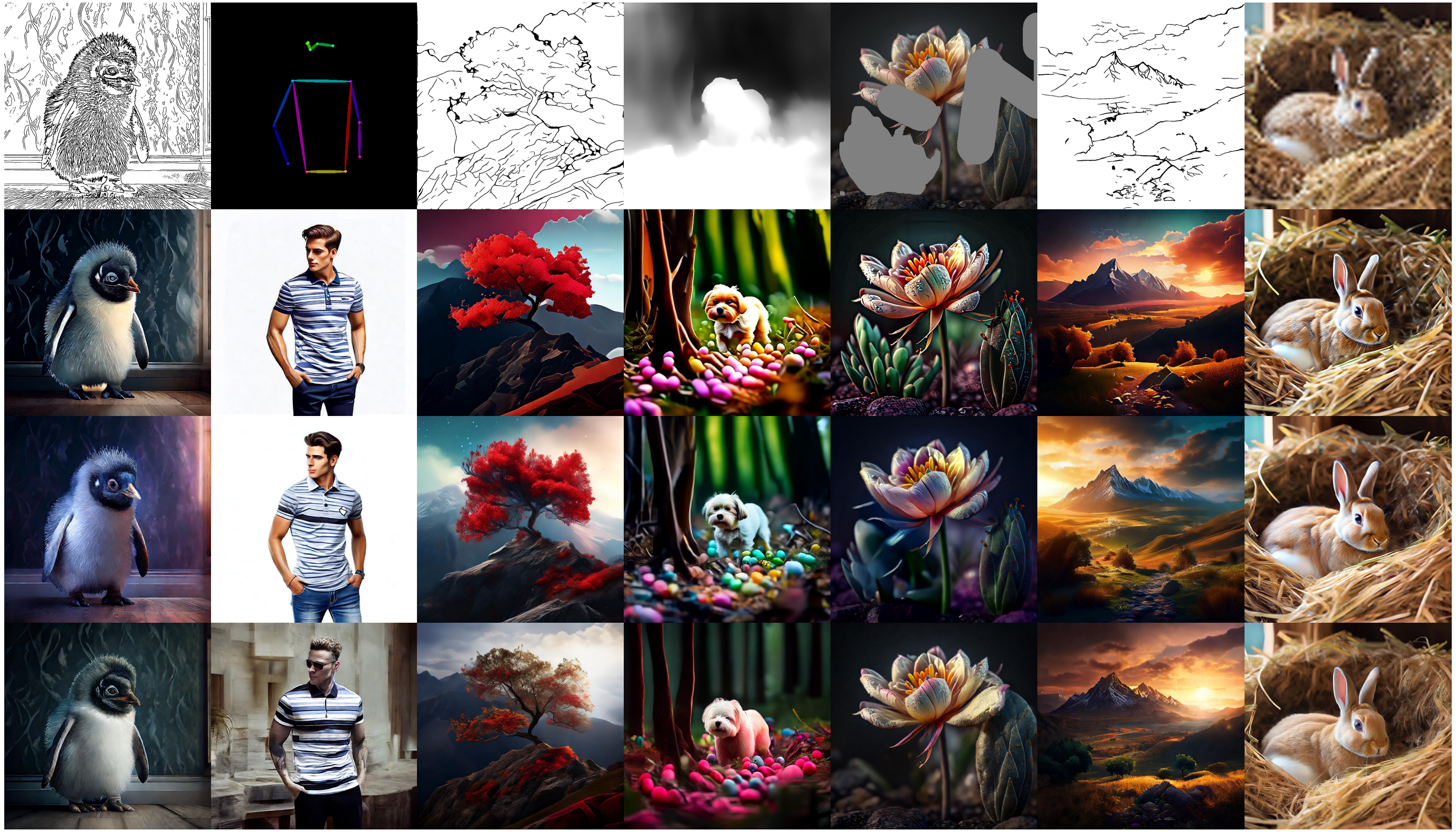
# Experiment

# Quantitative Results

| Task<br>Method | NFEs↓ | Time(s)↓ | Sketch2Image<br>FID↓/Fidelity↓ | Depth2Image<br>FID↓/Fidelity↓ | Mask2Image<br>FID↓/Fidelity↓ | 16×SR<br>FID↓/Fidelity↓ | Average<br>FID↓/Fidelity↓ |
|---|---|---|---|---|---|---|---|
| DM's ControlNet+DM | $50 \times 2$ | 23.6 | 8.40/0.106 | 11.48/0.177 | 4.37/0.085 | 5.01/0.121 | **7.31/0.122** |
| DM's ControlNet+CM | 1 | 0.2 | 30.71/0.083 | 26.08/0.193 | 14.67/0.431 | 21.32/0.237 | 23.19/0.231 |
| DM's ControlNet+CM+Adapter | 1 | 0.2 | 20.43/0.111 | 19.75/0.176 | 13.95/0.413 | 13.73/0.168 | 16.96/0.221 |
| CM's ControlNet+CM | 1 | 0.2 | 10.39/0.095 | 12.94/0.169 | 5.44/0.082 | 7.60/0.118 | **9.09/0.116** |
| DM's ControlNet+CM | 4 | 0.9 | 21.88/0.091 | 21.12/0.190 | 10.27/0.457 | 11.41/0.146 | 16.16/0.221 |
| DM's ControlNet+CM+Adapter | 4 | 1.0 | 11.91/0.113 | 12.83/0.175 | 9.16/0.452 | 7.21/0.146 | 10.27/0.221 |
| CM's ControlNet+CM | 4 | 0.9 | 9.30/0.103 | 9.87/0.175 | 4.98/0.110 | 6.31/0.134 | **7.61/0.130** |

Quantitative comparison of different methods. NFEs means the number of function evaluations. $\times 2$ for the diffusion model because classifier-free guidance is used. Time is recorded based on the generation of a $1024 \times 1024$ image.

# Visual Results

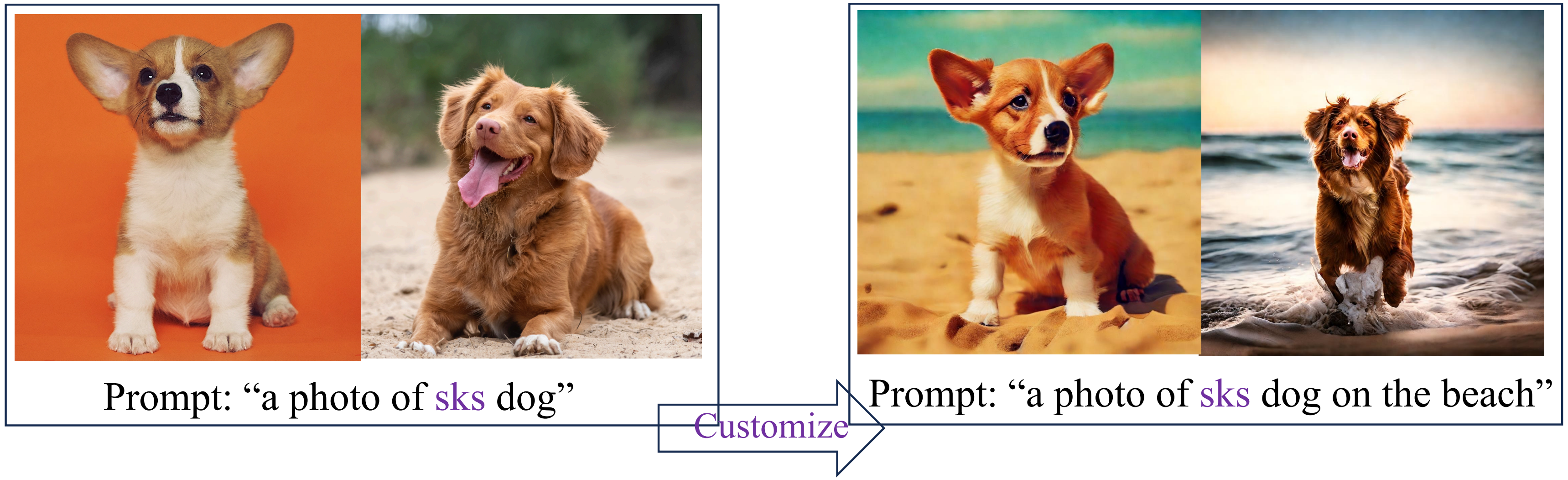



Visual comparison of different methods of adding controls.

# Visual Results

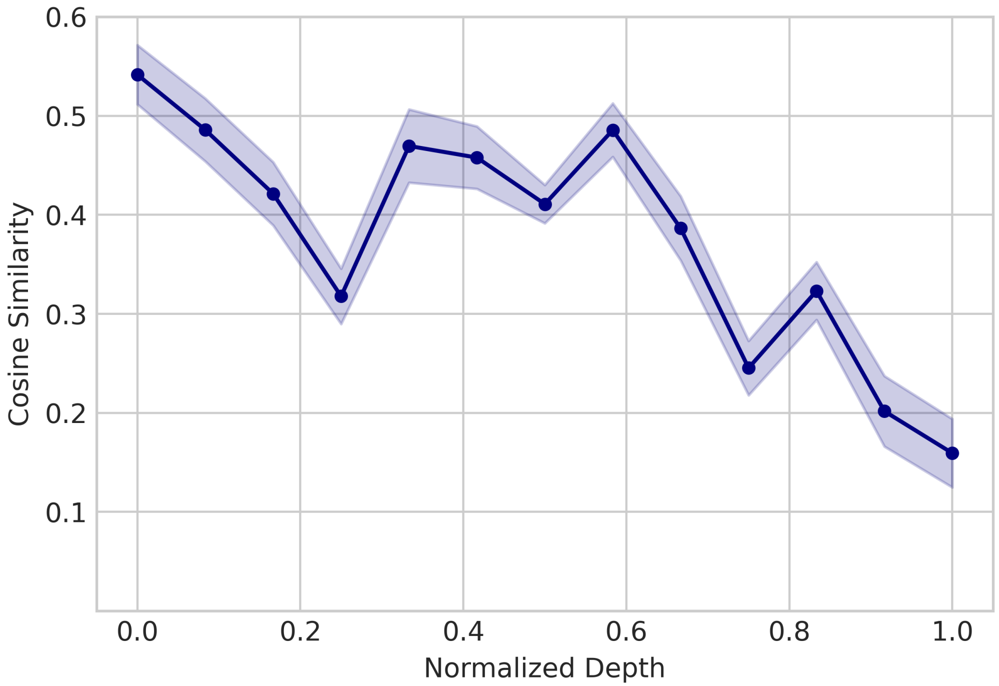
sketch | "a high-quality and professional image" | "A yellow dog lies on the grassland and enjoys the sun" | "watercolor style, a dog lies on the beach"

Visual results of CM's ControlNet with different prompts. Image resolution:1024×1024. NFEs: 4.

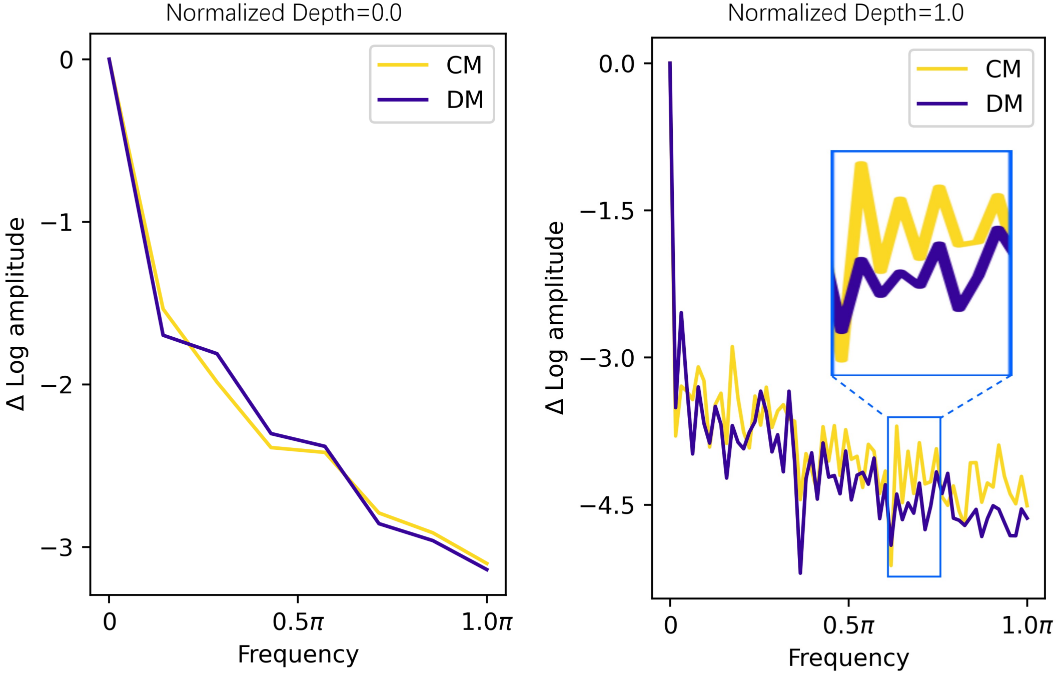



Visual results of customizing images using consistency training. Image resolution: 1024×1024. NFEs: 4.

# Analysis




(a) Cosine similarity across network depth between CM’s ControlNet and DM's ControlNet




(b) Log amplitude of Fourier-transformed control features from CM's and DM's ControlNet

Correlation analysis between CM's and DM's ControlNet. (a) shows the decreased correlation along the depth. (b) shows amplitude of Fourier-transformed features. These results validate that both ControlNets generally agree on high-level controls but differs on low-level controls.

# Conclusion

# Conclusion

- ControlNet of DM can transfer high-level semantic controls to CM; however, it often fails to accomplish low-level fine controls

- CM's ControlNet can be trained from scratch using the consistency training technique. Empirically, we can find that consistency training can accomplish more satisfactory conditional generation

- A unified adapter trained with the consistency training technique is capable of mitigating the discrepancy between DMs and CMs, thereby facilitating to transfer DM's ControlNet

# Thanks !