Tom & Jerry in Real life - Translating Cartoon to Natural Images using Stable Diffusion

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Sachin Salim, Shrikant Arvavasu, Nowrin Mohamed

Introduction

- Unpaired image-to-image translation (I2I) addresses challenges by transferring content between domains without explicit correspondences
- Various generative models like GANs, VAEs, and iterative models such as LDM excel in synthesizing realistic images
- Our specific focus involves translating Tom & Jerry images into realistic scenes, leveraging Stable Diffusion and BLIP to seamlessly facilitate content-rich adaptation

Data

- The collection of the dataset involved acquiring 5000 images for each of the source/cartoon and target/natural domains.
- Randomly selected 1200 images from the training set and used GPT-4 to generate image-action captions using OpenAI's APIs for fine-tuning BLIP.
- Datasets were sourced from Kaggle, Andresen et al., and Tom & Jerry show videos, each meticulously curated.

Methods

- Stable Diffusion adapted from Latent Diffusion Models (LDMs) facilitates high-quality generative modeling frameworks.
- LDMs employ a reverse diffusion process, allowing source-conditioned image generation for target domains.
- Overcoming challenges, the model incorporates class conditioning to align Tom with a cat and Jerry with a mouse.

$$z_t := \alpha_t z + \sigma_t \eta \qquad z = \mathcal{E}(x)$$

$$\hat{\theta} := \min_{\theta} \ \mathbb{E}_{z,c,\eta,t} \left[\|\hat{z}_\theta(z_t, t, c) - z\|^2 \right]$$

$$(x, y) \sim (\mathcal{X}, \mathcal{Y})$$

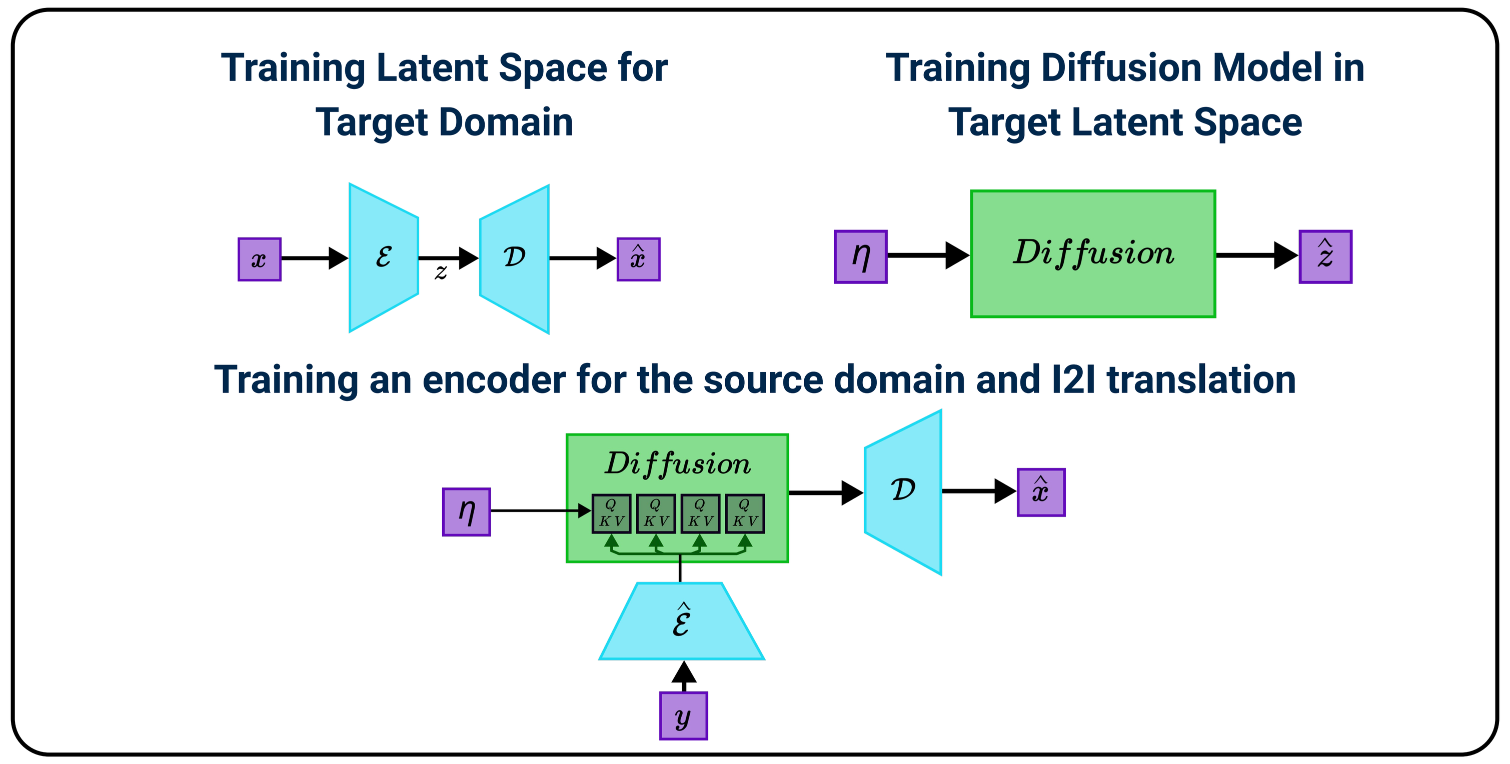
$$z \sim \mathcal{E}(x), \ y \sim \varphi(y)$$

$$Q = W_Q \cdot y, \ K = W_K \cdot z; V = W_V \cdot z$$

$$z_t^C = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

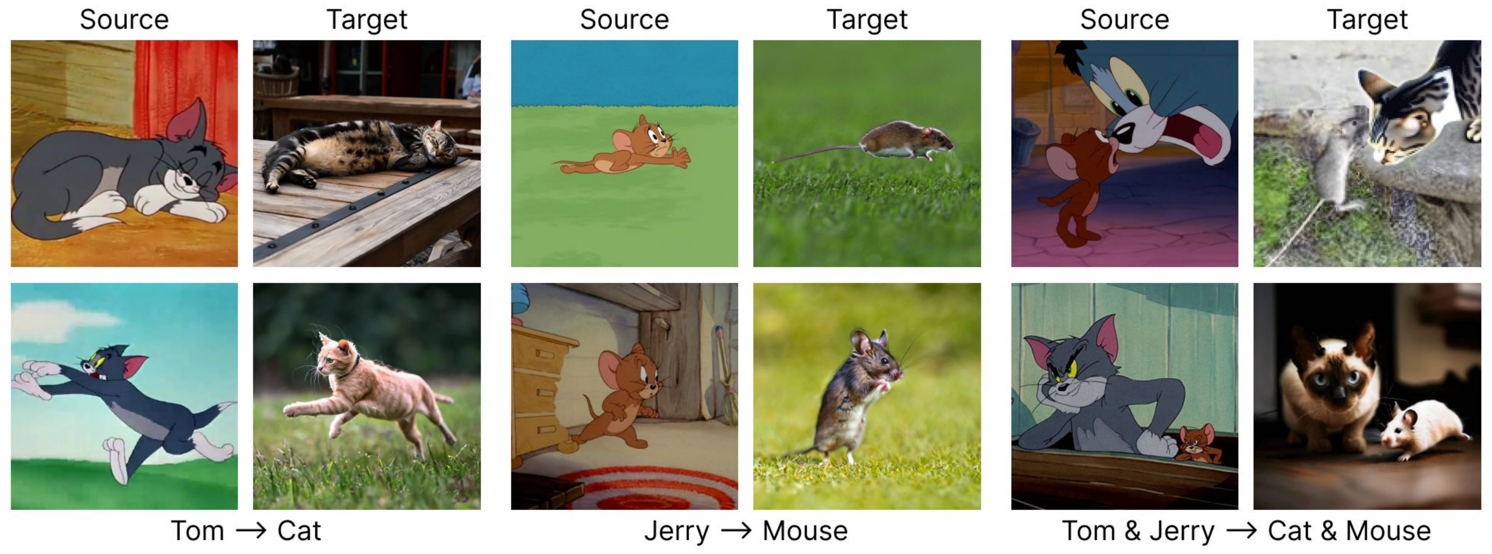
Methods

- Fine-tuned for scene description, BLIP, a multi-modal model, facilitates content transfer across image domains.
- Regularization with BLIP encoder similarity guides the diffusion model in content transfer between images.
- Project workflow: pre-train conditional diffusion, finetune BLIP for captioning, and enforce regularization for transfer.

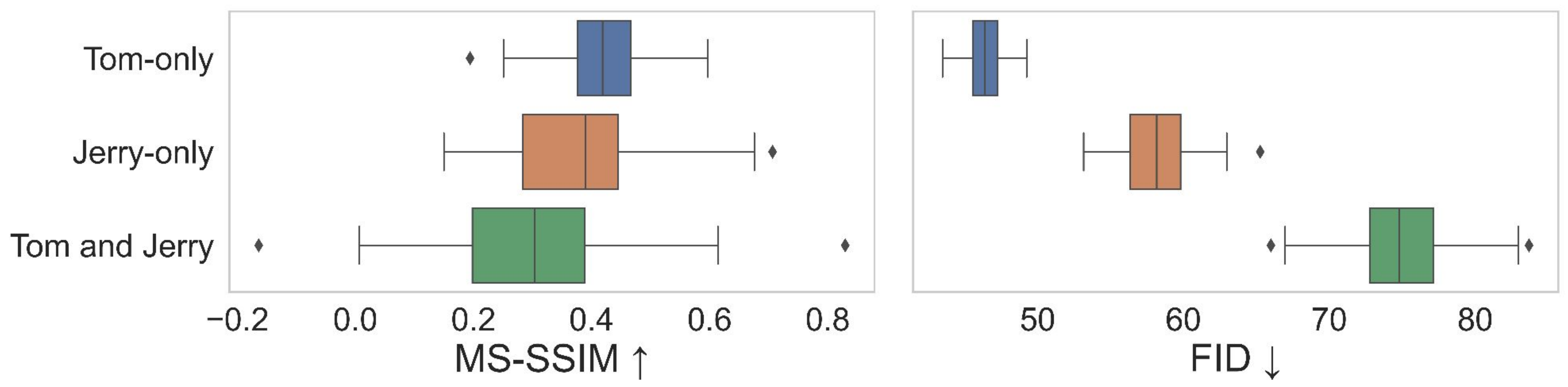


Results

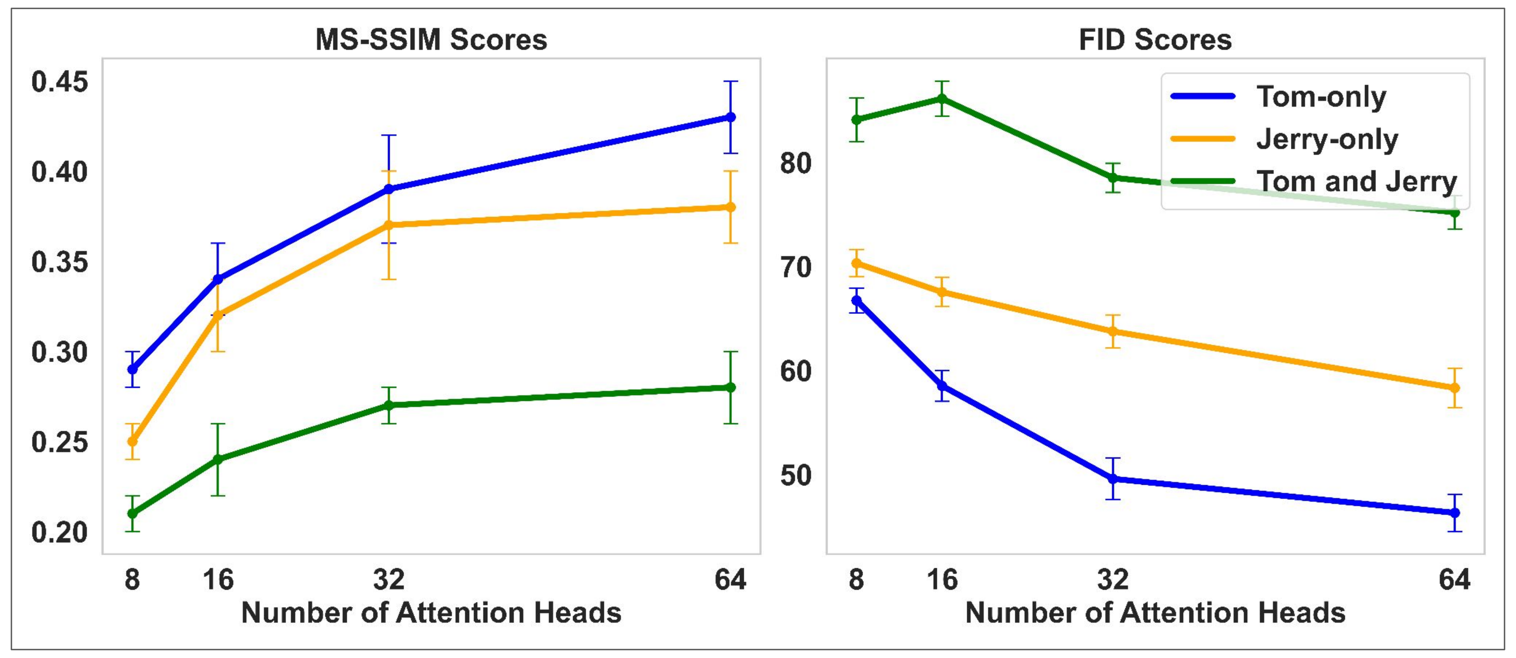
Comparison of Translation Results - Two visually optimal outcomes for each translation class



Quantitative results on generating natural images



Implementation Details



- Our project employed Python 3.9 and PyTorch framework for model training, utilizing an NVIDIA A40 GPU cluster.
- The latent diffusion model underwent a comprehensive 20-hours training period, encompassing both the latent space and the diffusion model itself.
- The source-conditioned generation model was trained for a focused duration of 8 hours.
- The BLIP model was trained for 15 hours to fine-tune its performance for highly descriptive captions

Conclusion and Discussion

1. Addressing fidelity challenges in cartoon translation may entail acquiring knowledge of segmentation maps to preserve action postures.
2. To improve LDM's slow inference, Markovian process could be modified akin to denoising diffusion implicit models.
3. Explore the extension of the model in video sequences. Investigate its performance in handling temporal variations and dynamic scenes, which is crucial for real-world applications.

Acknowledgements

Thank you, Dr. Jeong Joon Park, Dr. Stella Yu, Gaurav Kaul, and Zilin Wang for your exceptional guidance, support, and invaluable feedback throughout our course and project