Problem Statement & Datasets

Problem Statement:

Synthetic-to-Real Unsupervised Domain Adaptation for Image and Video Classification

Datasets:

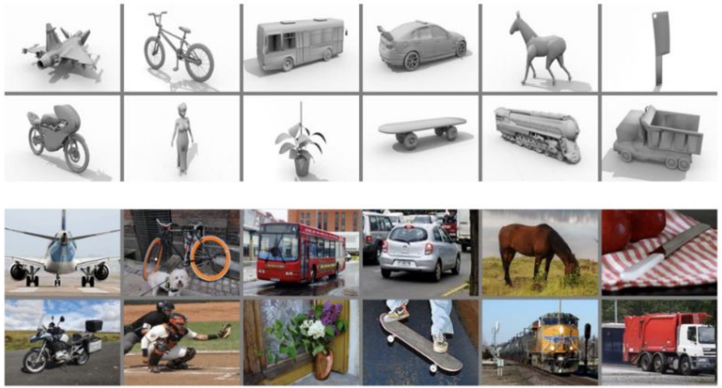
Images: VisDA-2017, S2RDA

Videos: RoCoG, Mixamo-Kinetics

Syn-to-Real Image Classification Datasets

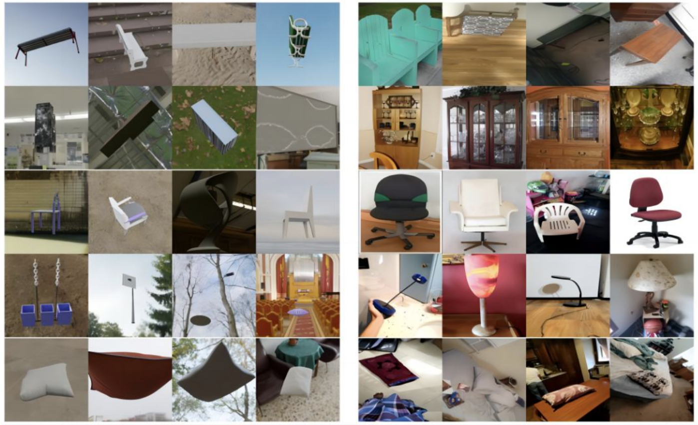
VisDA-2017

- 12 classes
- Source domain: 152397 images
  - Rendered 3D CAD models
- Target domain: 55388 images
  - Images from MS-COCO



S2RDA

- 49 classes
- Source domain: 588000 images
  - Rendered 3D models from ShapeNet
- Target domain: 60535 images
  - Imagenet, objectnet, visda, web



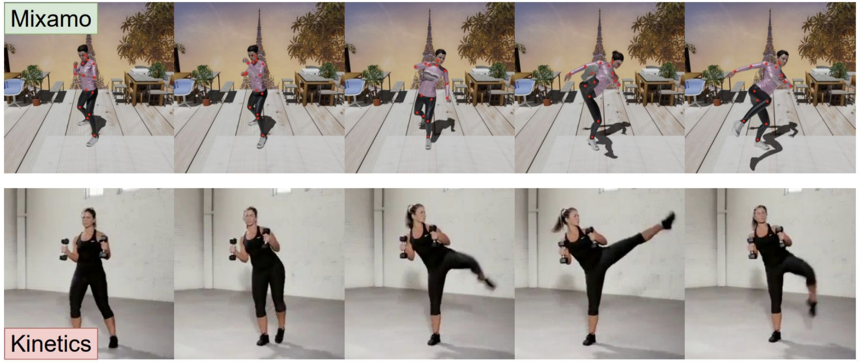
Syn-to-Real Action Recognition Datasets

RoCoG

- 7 classes
- US Army Field Manual

Mixamo-Kinetics

- 14 classes
- Subset of Kinetics dataset



Approach

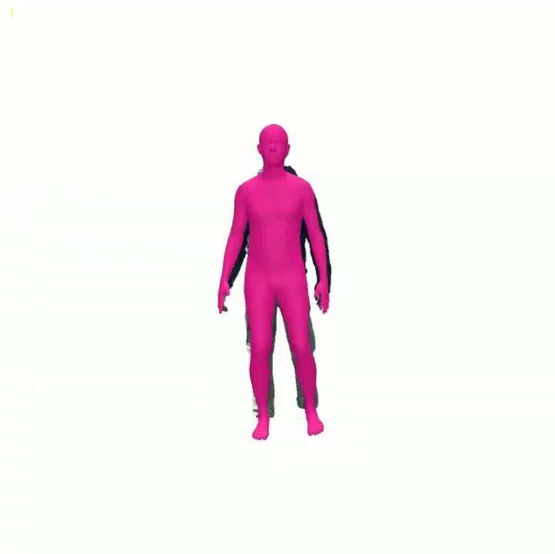
- Use diffusion-based image translation to reduce domain gap between real and synthetic data
- Transfer the style of target domain (real) to source domain (syn)

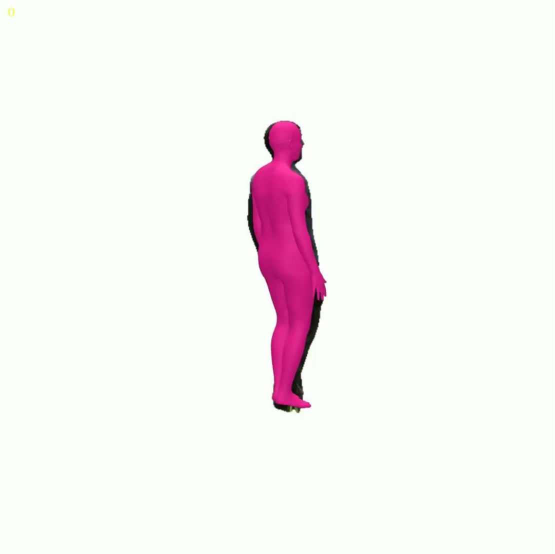
Motivation

- Prior work using conventional style transfer and GAN-based image translation for domain adaptation
- Diffusion-based style transfer methods demonstrate impressive results

4DH on RoCoG

Approach to reconstruct humans and track them over time.





Real video (Action: *"start"*)

Synthetic video (Action: *"Advance"*)

Experiment-1 – Source / Target Only

- Models: **Swin-Tiny** and **Swin-Base**
- They are trained, validated and tested on **VisDA-2017** dataset without domain adaptation
- Initialized using Imagenet pretrained weights
- Use results as lower/upper bound of performance

Experiment 1: Classification using Swin-T model (without domain adaptation)

- **Source only** classification with Swin-T backbone on VisDA

plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg (%)**
75	46	84	94	57	14	77	6	74	34	86	57	**60**

- **Target only** classification with Swin-T backbone on VisDA

plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg (%)**
91	79	85	97	82	69	80	83	95	57	85	90	**84**

Experiment 1: Classification using Swin-base model (without domain adaptation)

- **Source only** classification with Swin-B backbone on VisDA

plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg (%)**
85	40	81	97	58	16	80	10	77	56	83	49	**62**

- **Target only** classification with Swin-B backbone on VisDA

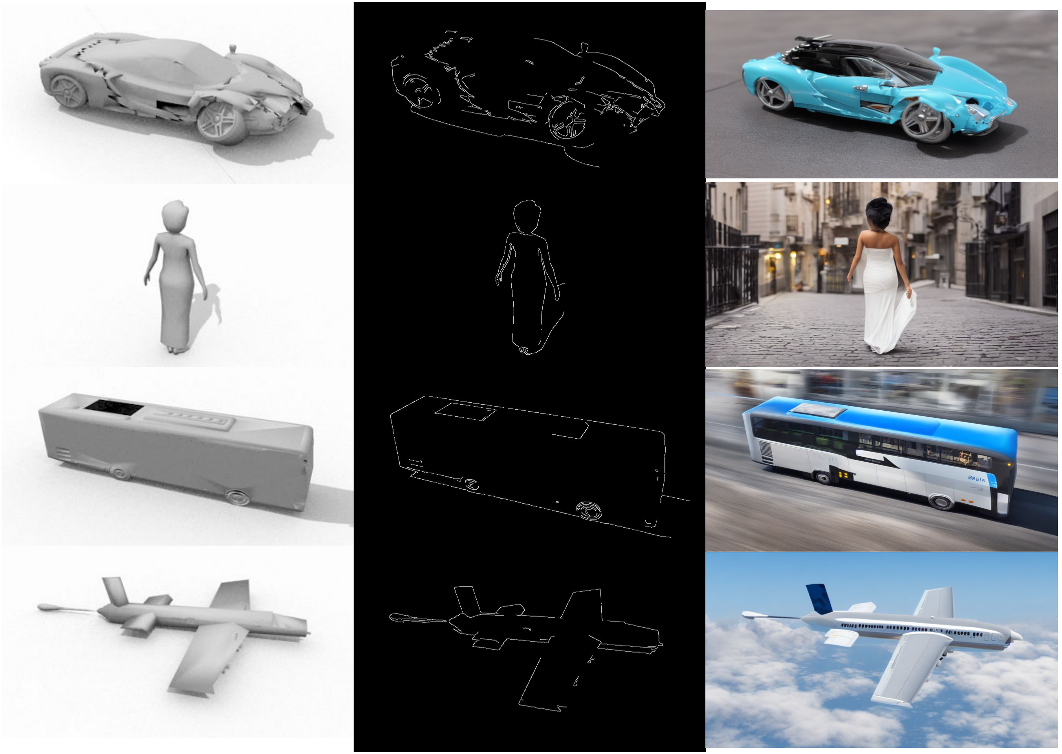
plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg (%)**
93	80	84	96	86	75	84	85	96	75	87	95	**87**

Experiment-2: Image Translation using ControlNet (without prompt engineering)

- Control Stable Diffusion 1.5 using ControlNet.
  - Using the **canny version** of ControlNet, obtained canny edge maps of synthetic train images.
  - These are used as a visual prompts for Stable Diffusion.
  - Text prompts used are in the format: **“*a {class_name}*”**.
  - Generated output: translated, "real” images with smaller domain gap with real target distribution
- Fine-tuned the previous Swin-base model on this “new”, translated training dataset for classification.
- Tested on the real target domain images of VisDA.

Experiment-2 Image Translation using ControlNET

- Using ControlNet conditioned on canny edges (without prompt engineering)



Experiment 2: Image Translation using ControlNet (without prompt engineering)

Classification using Swin Transformer model on translated VisDA images

- **Source only** acc (without DA): 60%
- **Target only** acc: 84%

- Using **Swin-tiny** model:

plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
89	69	85	91	70	55	76	46	85	32	74	80	**73**

- Using **Swin-base** model:

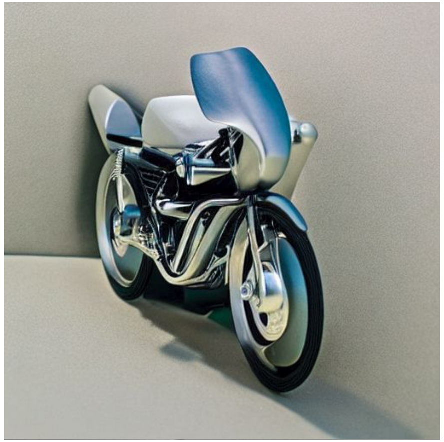
plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
93	70	88	95	77	51	79	53	85	28	78	73	**75**

- **Source only** acc (without DA): 62%
- **Target only** acc: 87%

Experiment 3: Translate synthetic VisDA images with "target prompts"

- Used samples from VisDA validation set
- Passed through **BLIP** captioning model to generate captions
- Then used these as "*target prompts*" for translation with canny version of ControlNet.

Generated image



Without "target prompts"

Generated image

With "target prompts"

Target set image

Target prompt: "*a motorcycle parked on the side of a street*"

Experiment 3: Classification using Swin Transformer model on translated VisDA images with "target prompts"

- **Source only** acc (without DA): 62%
- **Target only** acc: 87%

Translation without target prompts: 75.14%

- Using Swin-base model:

plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg (%)**
93	52	88	94	74	55	81	49	86	42	80	83	**76.76**

- Val acc increases from
  - 69.79% (without target prompts) to
  - 74.52% (with target prompts)

Experiment-4: Capture target style using Textual Inversion

- Adding a style-based token in the embedding space of SD.
- Fine-tune the embedding of SD to create personalized images based on custom style. Instead of re-training the model, we can represent the custom style as new words in the embedding space of the model. As a result, the new word will guide the creation of new images in an intuitive way.
- Textual Inversion creates an additional "word" that is added to the base model's vocabulary so it can draw it.
- Embeddings are smart compressions of data (images, text, audio, etc) into numerical representations.
- By training new embeddings for Stable Diffusion, we can give it a new point to try to get close to as it removes noise.

Experiment-4: Method

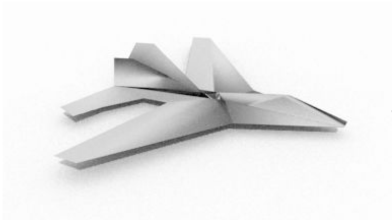
- Using a few images from all classes of VisDA val set, we trained a new embedding for Stable Diffusion.
    - *initializer_word* = [“realism”, “style”]
    - *placeholder_token*= “*”
- Loaded this new generated embedding (describing the style of real domain) in the Stable Diffusion ControlNet Pipeline.

Experiment-4: Method

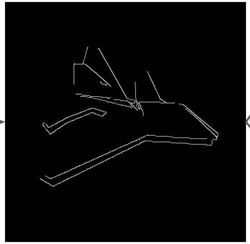
BLIP captioning model

“a small plane sitting on top of a runway”

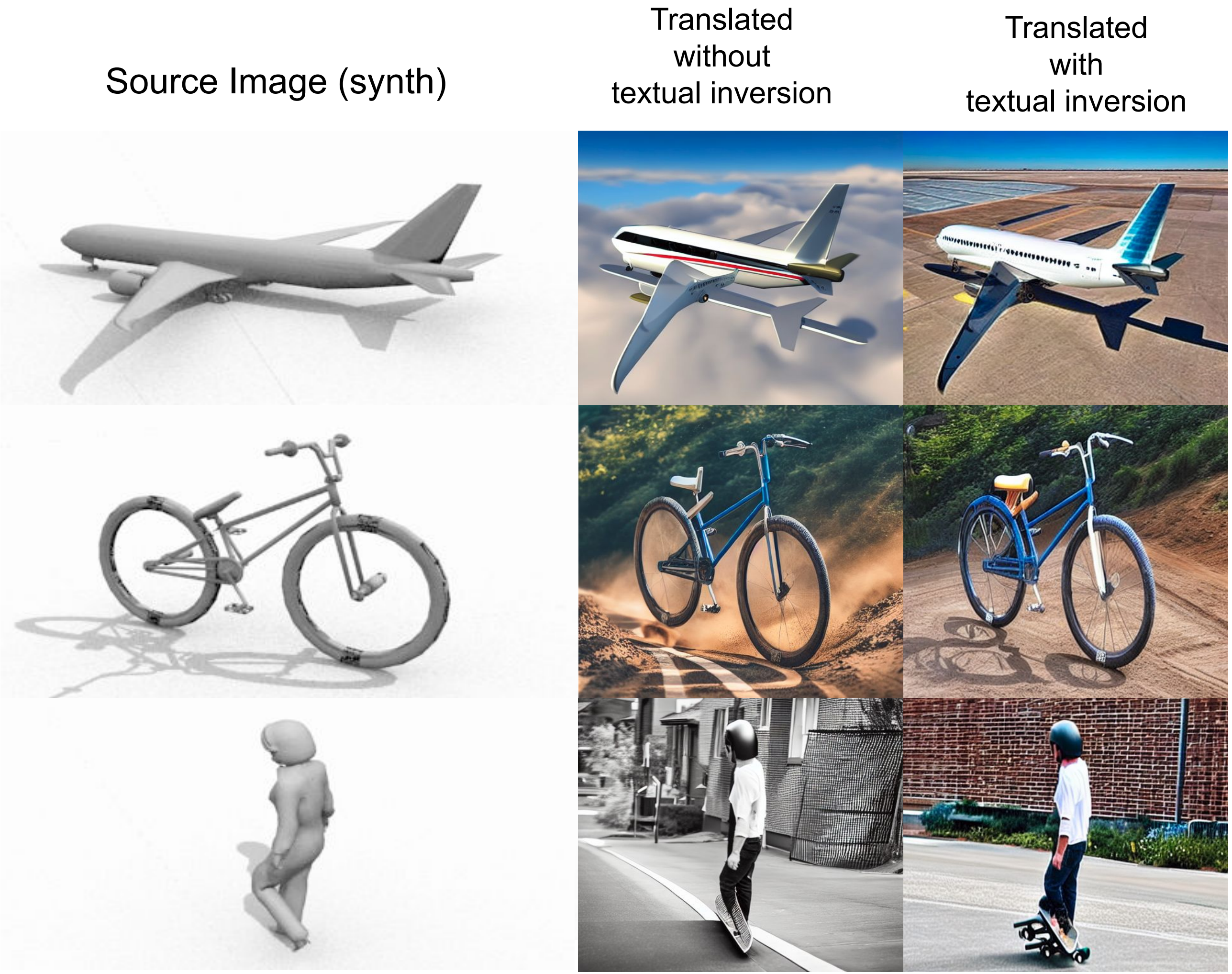
New learnt embedding with token ‘*’ capturing the style of real domain images



Canny edge map generation



*“a small plane sitting on top of a runway in the style of *”*


Source Image (synth)
Translated without textual inversion
Translated with textual inversion

Experiment 5: Classification using Swin-Transformer on translated VisDA images with “target prompts” and Textual Inversion

- **Source only** acc (without DA): 62%
- **Target only** acc: 87%

Translation without target prompts: 75.14%

Translation with target prompts: 76.76%

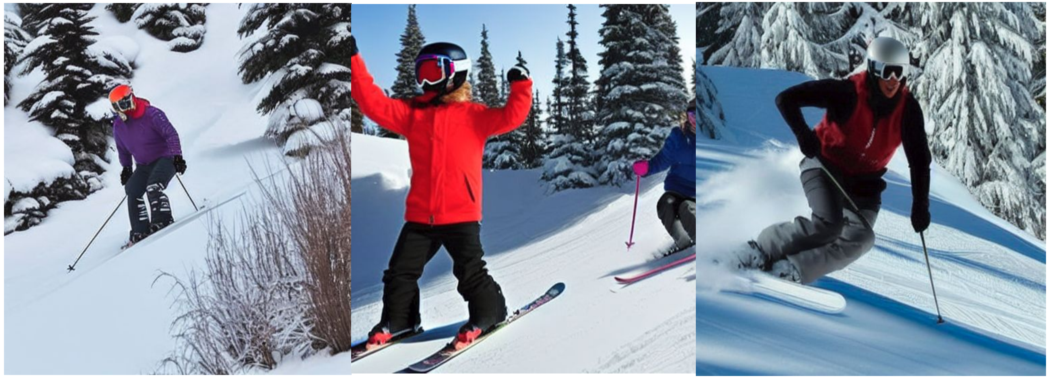
- Using Swin-base model:

plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg (%)**
91	59	89	96	82	50	87	47	91	49	73	84	**76.78**

- Val acc increases from
  - 74.52% (with target prompts) to
  - 76.51% (with target prompts+TI)

Experiment-6: Capture target style using Dreambooth

- Used all the val set images to fine-tune stable diffusion v1-5 model.
- Tested this model with a few *“target prompts”.* An example:
  - Prompt: *“a person skiing in the style of sks object”*. (sks is the placeholder token/identifier)
  - Generated Images:



Experiment-7: Translation of VisDA synthetic images using CN+dreambooth

- Translated VisDA synth images with new, fine-tuned “personalized” SD v1-5 model without captions or TI embedding.
- Prompts used: “a {*class_name*} in the style of sks object”.
- Some generated examples:

Experiment-8: Translation of VisDA synthetic images using CN+TI+dreambooth

- Generated new TI embedding by training it on all val images for the personalized model as base model.
- Loaded new embedding into the embedding space of the personalized SD v1-5 model.
- Translated visDA synth images with personalized stable diffusion model with new textual inversion token for more conditioning.
- Some generated examples:
    - Prompt: “a {*class_name*} in the style of sks object, in the style of valid”
    - Generated images:

Experiment 9: Classification using Swin-Transformer on translated VisDA images with CN+TI+dreambooth

- **Source only** acc (without DA): 62%
- **Target only** acc: 87%

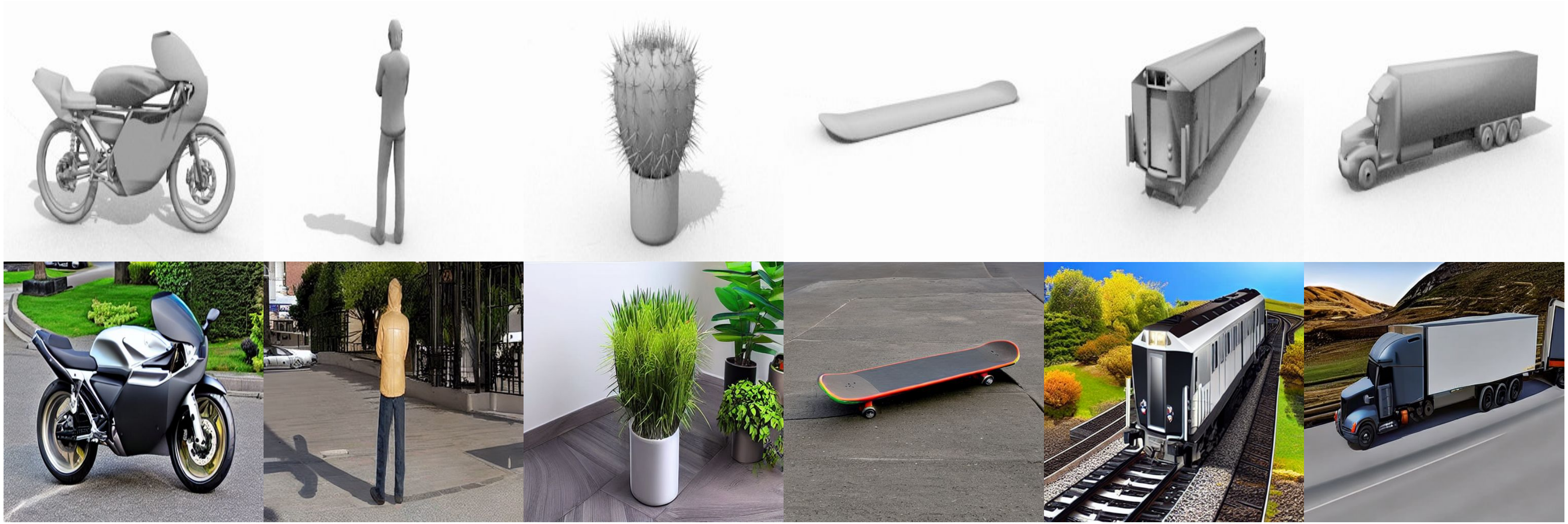
Translation without target prompts: 75.14%

Translation with target prompts: 76.76%

Translation with target prompts+TI: 76.78%

- Using Swin-base model:

plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg (%)**
93	64	86	95	88	58	87	46	93	29	72	84	**77.2**



Additional Steps

- Hyperparameter tuning in classification task.
- Varying guidance scale to incorporate variation in training Stable Diffusion.
- Running Text-to-image generation using: basic label, hard engineered, language enhanced prompts (T5-base fine-tuned on CommonGen).

Crafting Prompts

Number	Variations
(1)	high quality, low quality, blurred, bad, old, closeup
(2)	person, car, truck, bus, bicycle, motorcycle
(3)	on a street, on a road, in a forest, in a city, on the sidewalk, on the highway

Label	HE	LE
A photo of a person	A (1) photo of a (2) (3)	person with some kind of cd and
A photo of a car	A (1) photo of a (2) (3)	car coming out of the garage
A photo of a truck	A (1) photo of a (2) (3)	several trucks are out in the wind
A photo of a bus	A (1) photo of a (2) (3)	A school bus in a rural area travelling
A photo of a motorcycle	A (1) photo of a (2) (3)	woman riding a motorcycle with her
A photo of a bicycle	A (1) photo of a (2) (3)	A group of bicycles is in the shop

Experiment 10: Translation of RoCog Synthetic frames using MultiControlNet + Dreambooth

- We can effortlessly combine ControlNet with fine-tuning. For example, we can fine-tune a model with DreamBooth, and use it to render images into different scenes.
- Fine-tuned stable diffusion v1-5 model on the real frames of RoCoG and used this personalized model in the StableDiffusionControlNet pipeline.
- Used a MultiControlNet model with two conditionings of: edge and pose.
  - Prompt used: “a person in the style of sks person” —> ‘sks’ is the placeholder token (identifier).
  - Some Results:

Synthetic Frame

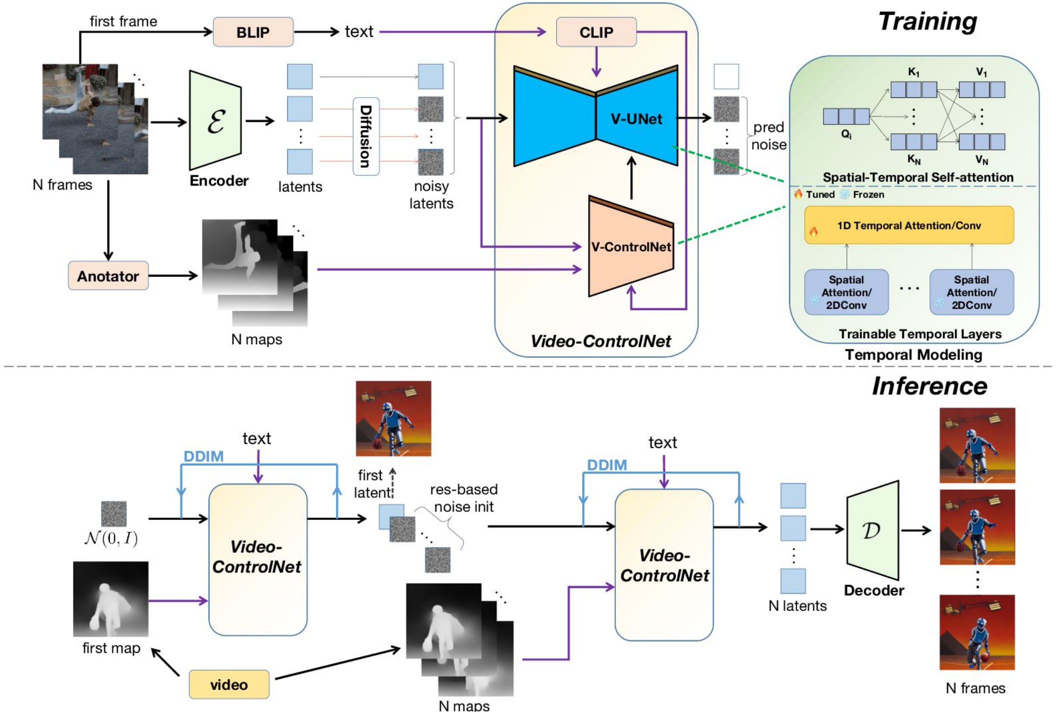
Translated Frame

Real frame sample used for fine-tuning SD

Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models

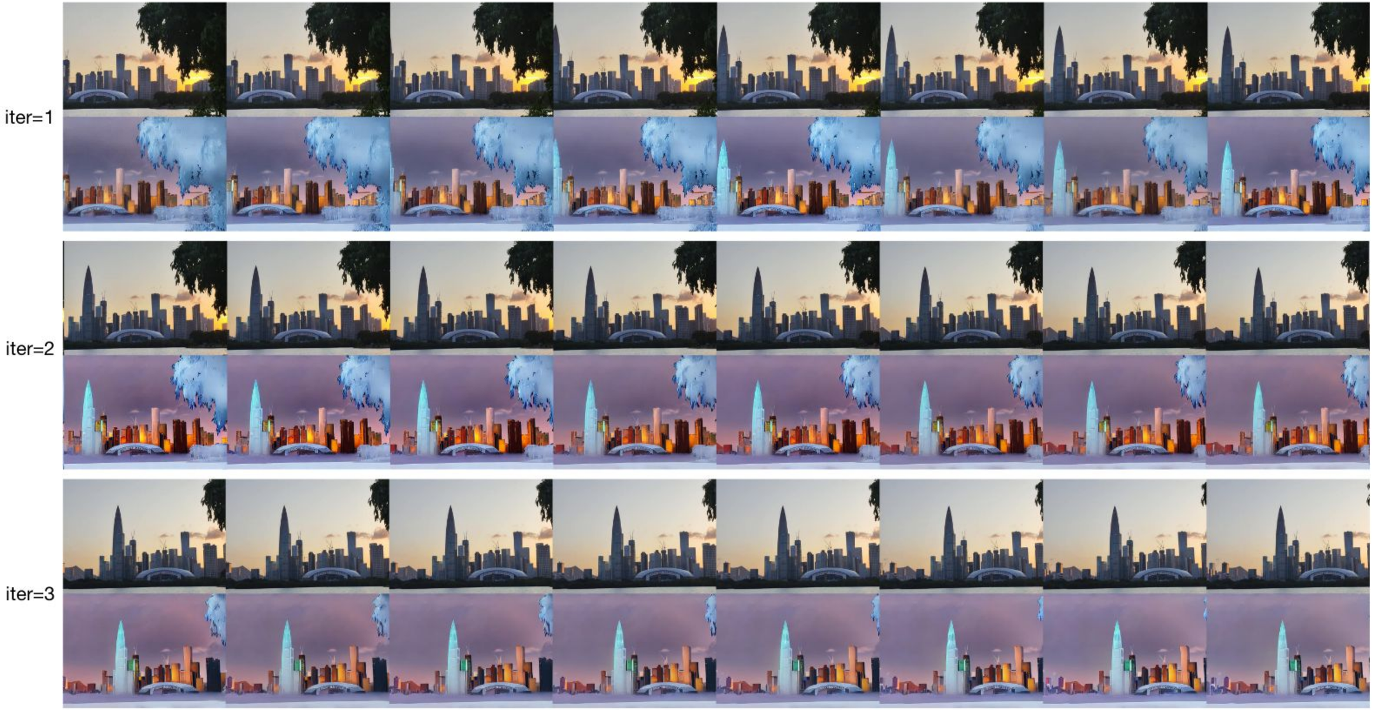
- A controllable T2V framework that is capable of generating videos conditioned on text prompts and control maps.
- Introduce a residual-based noise initialization strategy that incorporates motion from the input video into the diffusion process, resulting in the generation of videos that are less flickering and motion-aligned.
- Present a novel first-frame conditioning strategy that not only empowers the model to generate videos generalized from the image domain but also to generate arbitrary-length videos auto-regressively.
- Experiments demonstrate that this framework is capable of generating higher-quality, more consistent videos using fewer training resources.

Method



first frame
BLIP
text
CLIP
Training
N frames
Encoder
latents
Diffusion
noisy latents
V-UNet
pred noise
V-ControlNet
Anotator
N maps
Video-ControlNet
Spatial-Temporal Self-attention
Tuned
Frozen
1D Temporal Attention/Conv
Spatial Attention/ 2DConv
Trainable Temporal Layers
Temporal Modeling
Inference
DDIM
text
$\mathcal{N}(0, I)$
Video-ControlNet
first latent
res-based noise init
first map
video
N maps
N latents
Decoder
N frames

Results



frozen city, high-quality, realistic.