Group 18 Project Presentation

313552049 鄭博涵 312552026 蔡濟謙 110550035 陳奎元

Outline & Introduction

- Background: SG-I2V
- Failed Attempt: Modernize
- Experiment: Analyze
- Conclusion

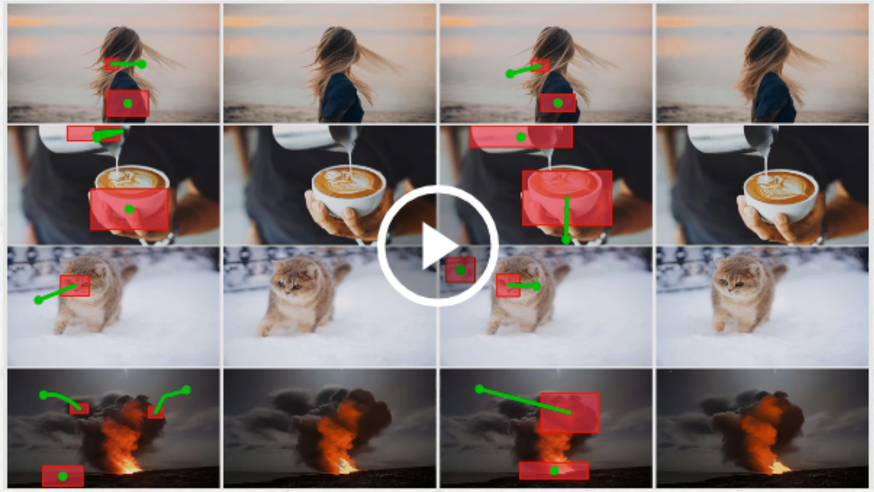
Background

Basis: SG-I2V: SELF-GUIDED TRAJECTORY CONTROL IN IMAGE-TO-VIDEO GENERATION

- Author: Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, David B. Lindell
- Source: ICLR 2025 submission 3365 at 24 Sept 2024

What is SG-I2V?

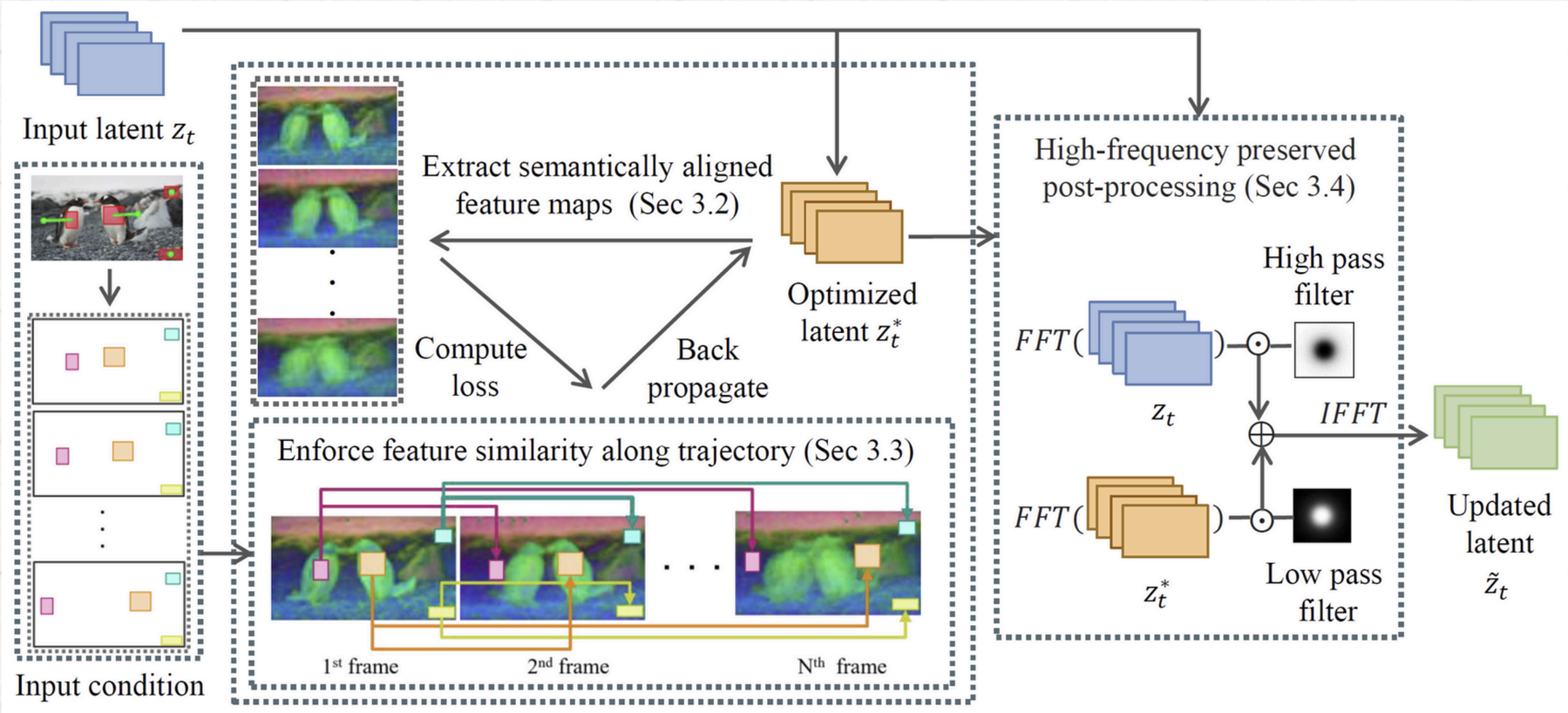
- Image-to-video generation framework
- Zero-shot trajectory control



Why do we need SG-I2V?

- Many practical applications
- Avoid cost of fine-tuning
- No similar framework available

Background


Input latent z_t
Input condition
Extract semantically aligned feature maps (Sec 3.2)
Compute loss
Back propagate
Optimized latent z_t^*
Enforce feature similarity along trajectory (Sec 3.3)
1st frame
2nd frame
Nth frame
High-frequency preserved post-processing (Sec 3.4)
High pass filter
$FFT($
z_t
$IFFT$
$FFT($
z_t^*
Low pass filter
Updated latent $\tilde{z}_t$

Modifications and Motivations

- Modify the structure of the U-Net with a diffusion model in SG-I2V to a DiT model structure (Transformer Diffusion).

- Because U-Net model is very popular and little traditional, so I think if we modify it to DiT model.
  The performance will be better or not?

Structure of DiT model

The DiT Block is use:

- Layer Normalization
- Self-Attention Mechanism
- Cross-Attention
- Feed-Forward network (MLP)
- Residual Connections

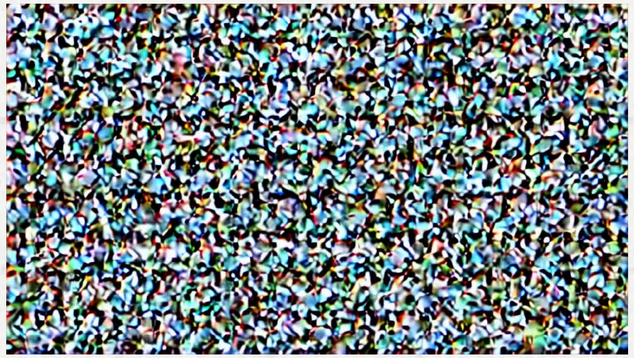
```python
class DiTBlock(nn.Module):
    def __init__(self, hidden_dim, num_heads, mlp_ratio, use_cross_attention=True):
        super().__init__()
        self.attn_norm = nn.LayerNorm(hidden_dim)
        self.mlp_norm = nn.LayerNorm(hidden_dim)
        self.multi_attn = nn.MultiheadAttention(hidden_dim, num_heads, batch_first=True)

        self.use_cross_attention = use_cross_attention
        if use_cross_attention:
            self.cross_attn_norm = nn.LayerNorm(hidden_dim)  # let normalize input data before attn_layer
            self.cross_attn = Attention(
                query_dim=hidden_dim,
                cross_attention_dim=hidden_dim,
                heads=num_heads,
                dim_head=hidden_dim // num_heads,
                dropout=0.0,
                bias=False,
                only_cross_attention=False,
                # added_kv_proj_dim=hidden_dim
            )

        # MLP (Feed-Forward network)
        self.mlp = nn.Sequential(
            nn.Linear(hidden_dim, int(hidden_dim * mlp_ratio)),
            nn.GELU(),
            nn.Linear(int(hidden_dim * mlp_ratio), hidden_dim)
        )
```

When integrated into the DiT model, these blocks are arranged sequentially to build a hierarchical transformer

Result



Since we found that there isn’t an adequate pre-trained model available, most pre-trained DiT models are designed for generating static images rather than dynamic videos.

As a result, our attempt failed. While the modified code is executable, it only produces noisy videos, as shown on the top-right.

In conclusion, we believe that modifying the DiT model for this purpose is infeasible in short time.

Therefore, we decided to shift our focus to a different task:
Analyzing and generating videos for each category (e.g. human, animal).
And we plan to use new example data to execute the original SG-I2V architecture.

Experiments

Analysis SG-I2V performance.

20 photos related to people, animals, and vehicles.

Task 1: The quality of generated images across different types of photos.

Task 2: The impact of video length on generation quality.

Task 3: Whether objects move in the correct direction.

Task 1

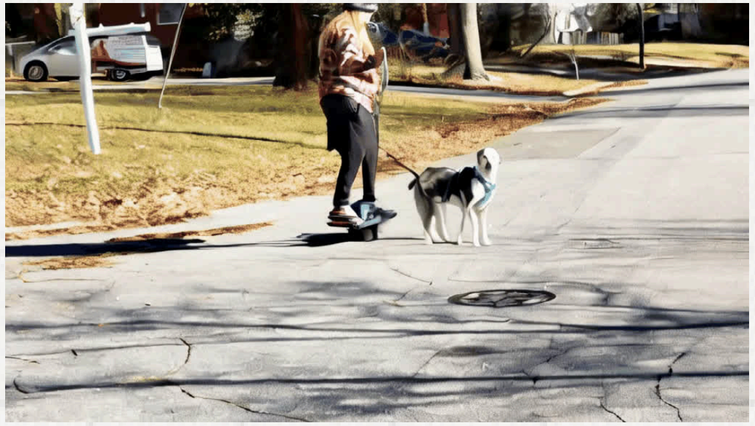
The Quality of Generated Videos.

Category	Good	Average	Poor
People	6	10	4
Animals	9	6	5
Vehicles	13	3	4

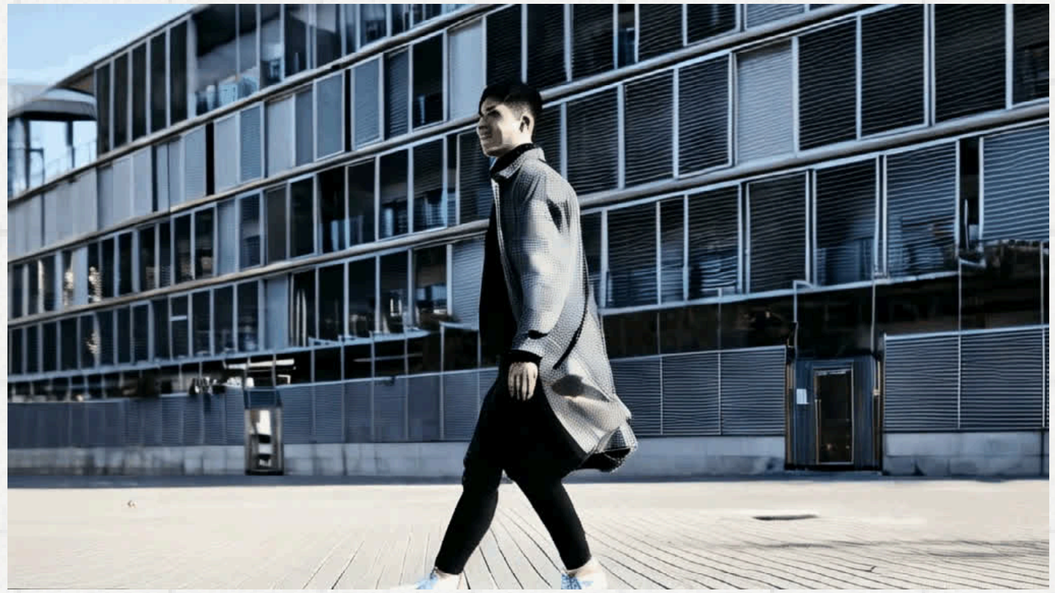
Good

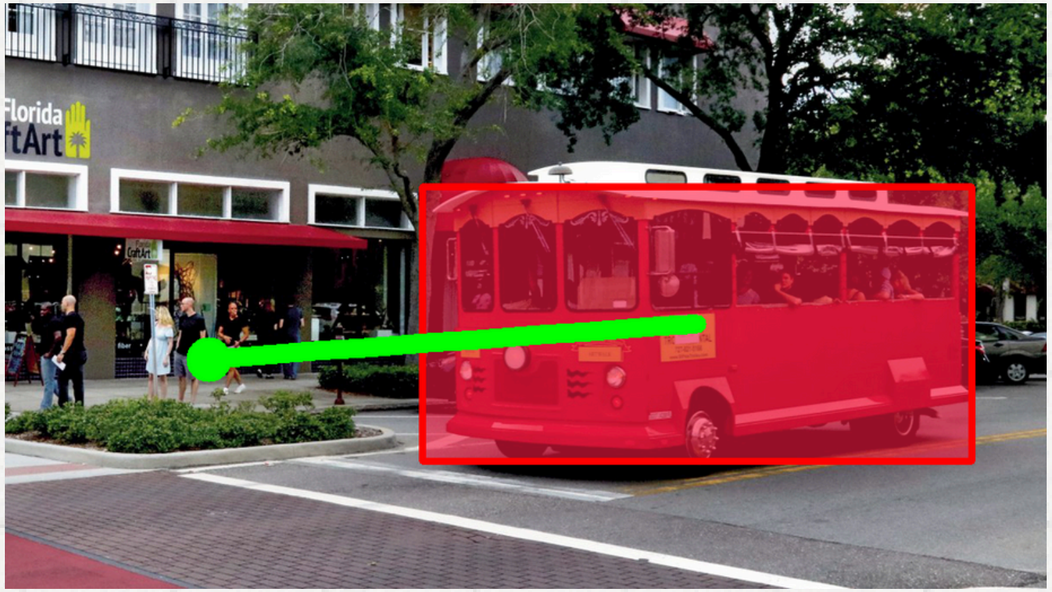
Average

Poor

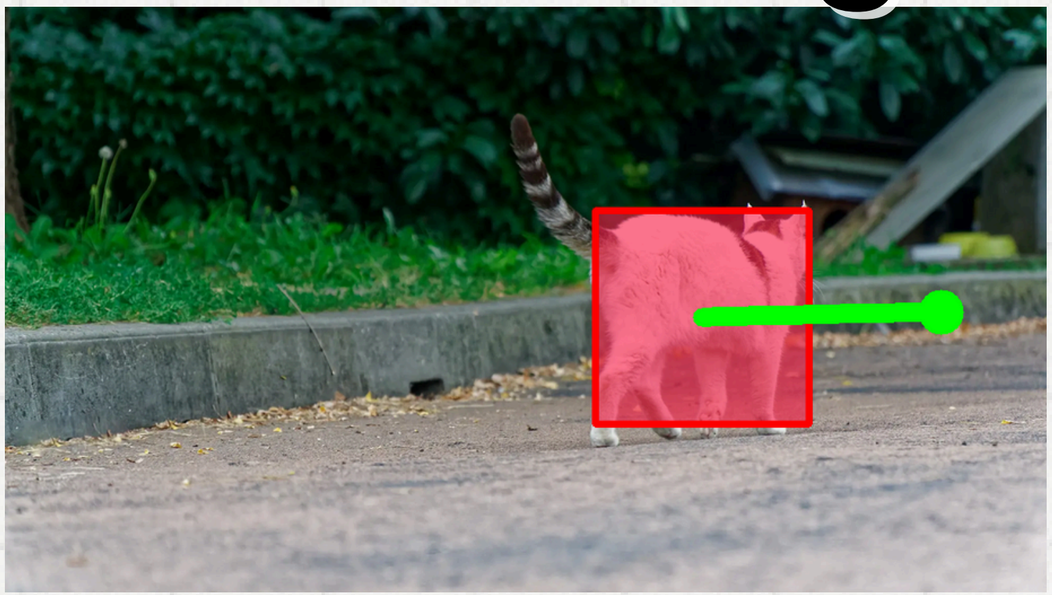


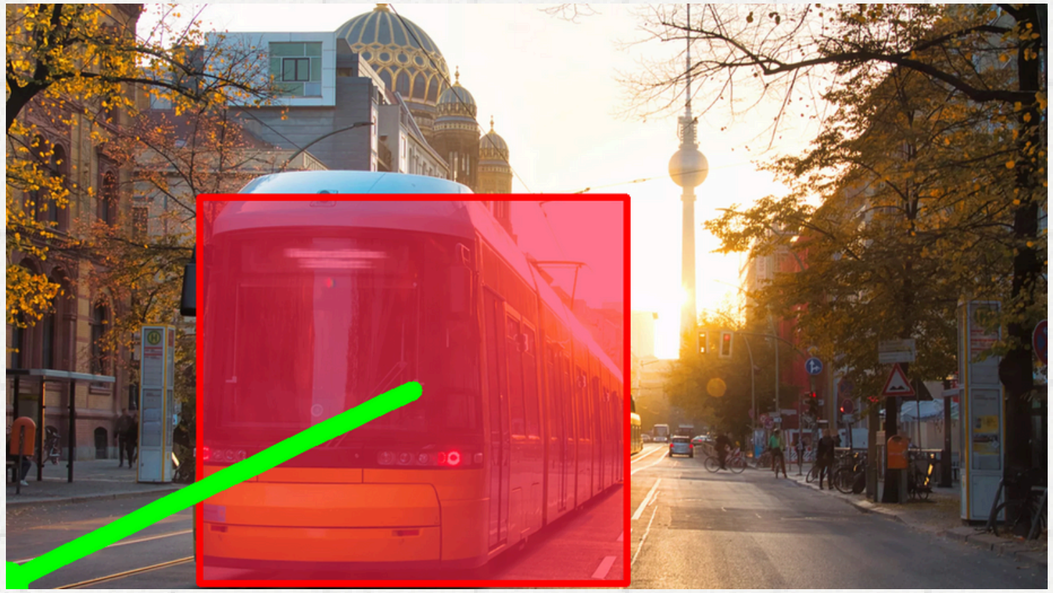
Task 1: Good Results



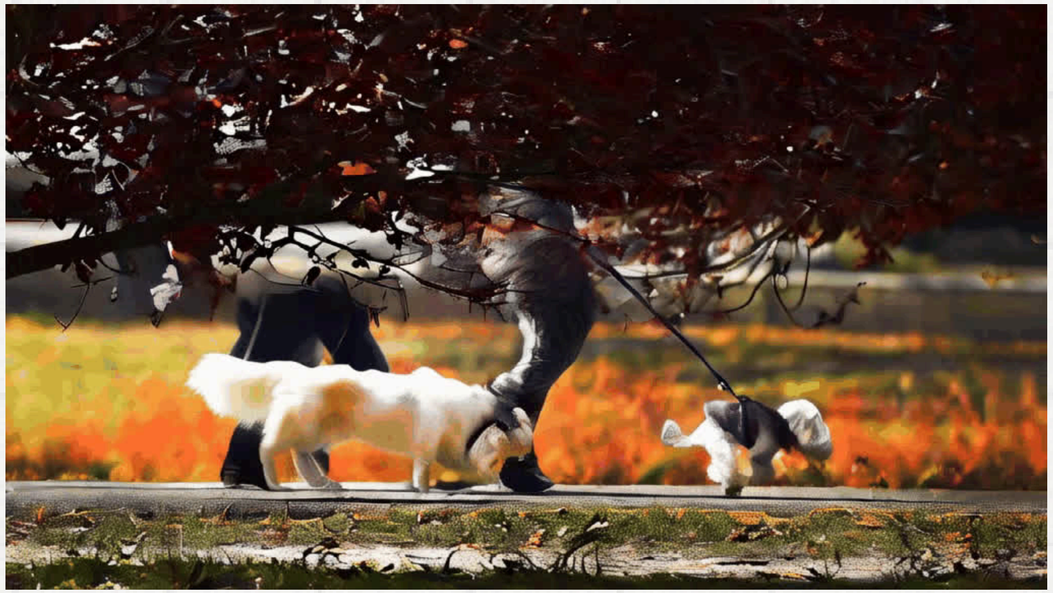


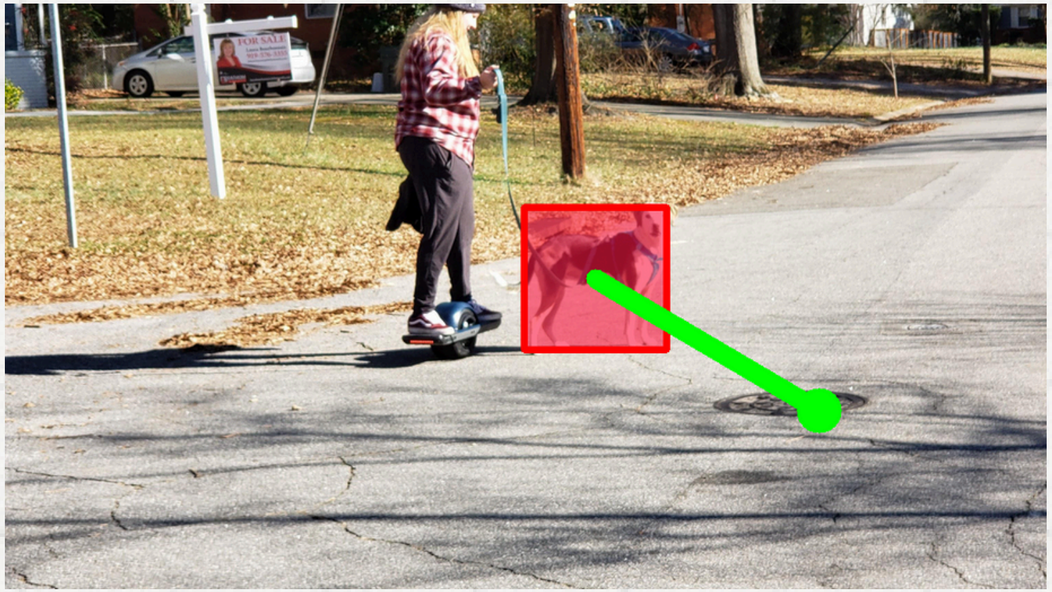
Task 1: Average Results

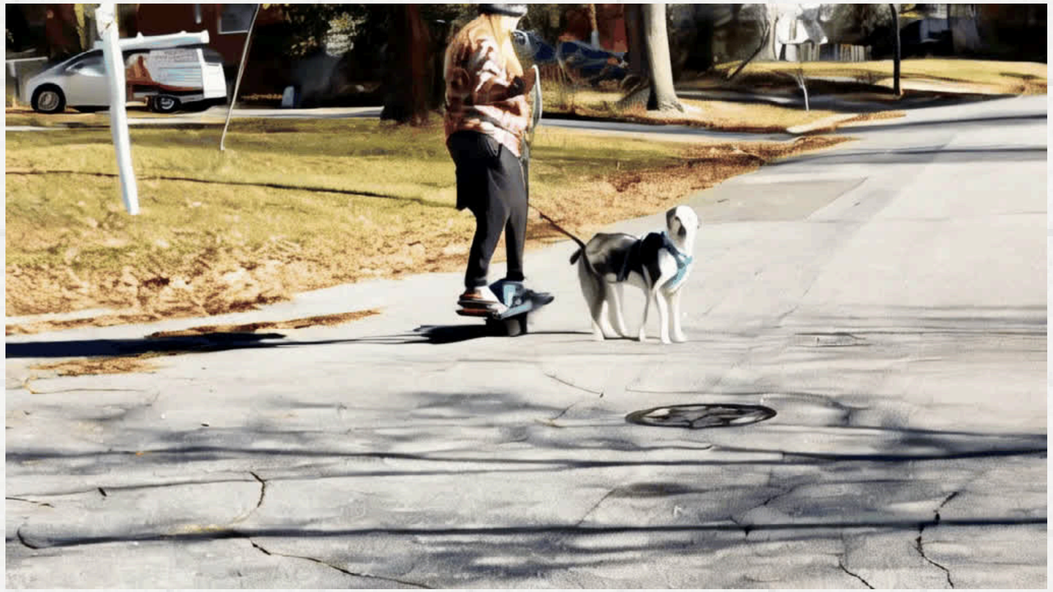




Task 1: Bad Results

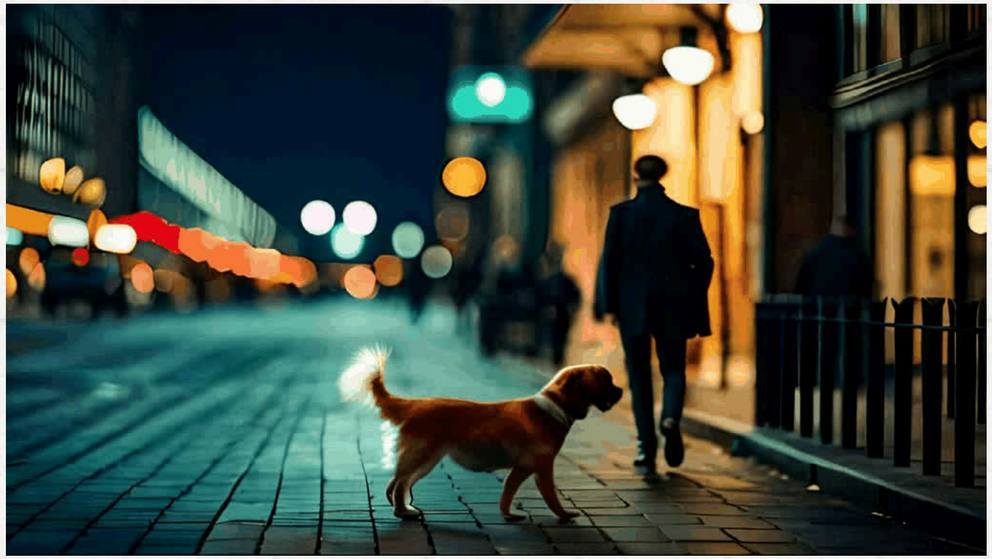


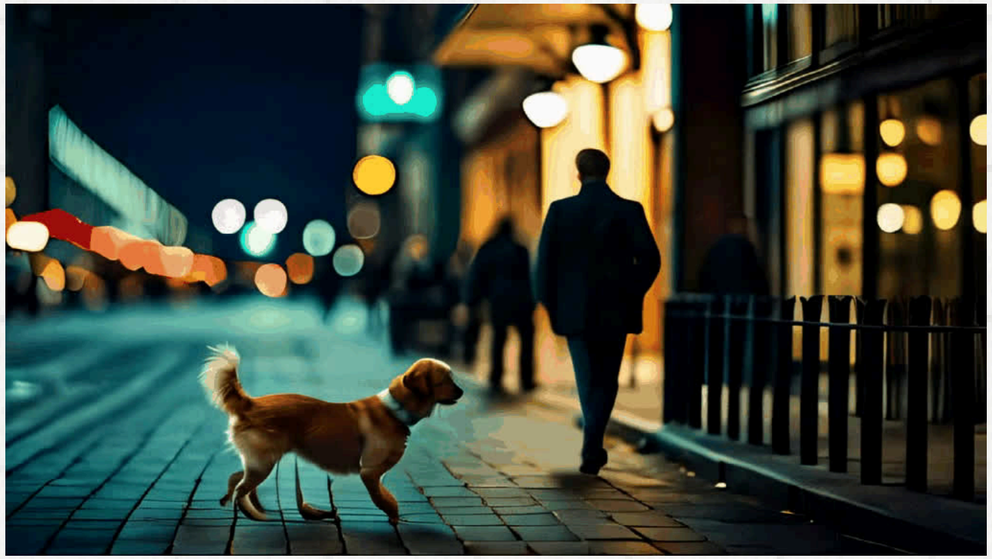




Task 2

Different length of the video generation





7

14

Task 2

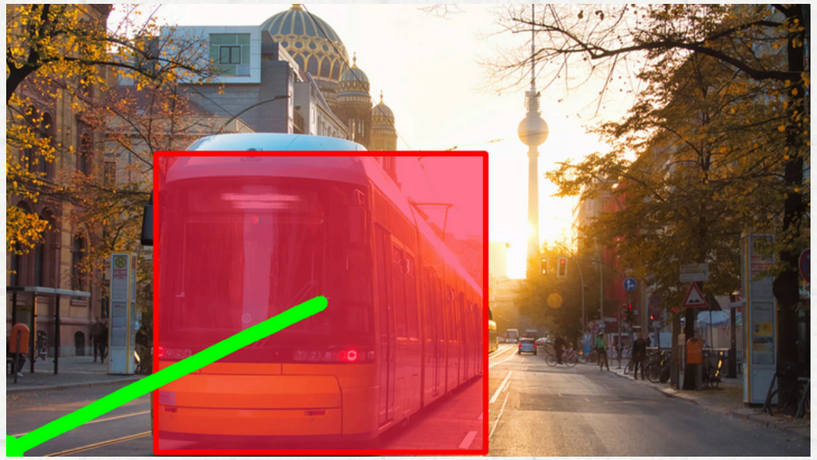
Is the camera fixed or moving?

Length of Frames	7		14	
	Fixed	**Moving**	**Fixed**	**Moving**
People	11	9	5	15
Animals	11	9	1	19
Vehicles	12	8	7	13

Task 3

Is it moving in the right direction?

Length of Frames	7		14	
	Wrong	**Correct**	**Wrong**	**Correct**
People	1	19	2	18
Animals	4	16	2	18
Vehicles	0	20	4	16



Conclusion

- SG-I2V: first framework for zero-shot trajectory control in image-to-video generation.
- Analyzes diffusion features, showing pre-trained models effectively guide motion generation.
- Demonstrates effectiveness through quantitative and qualitative results on synthetic and real-world images.
- Unveils inner mechanisms of diffusion models to inspire future designs.
- Offers a novel approach to improve video generation via trajectory control.

Thank you for listening!