ICLR

SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan,
Zuozhu Liu, Haoji Hu, Pengfei Wan, Di Zhang

jianhongbai@zju.edu.cn

Zhejiang University

KwaiVGI Lab

April, 2025

Jianhong Bai

Overview

TL; DR:

SynCamMaster generates multiple synchronized videos of the same dynamic scene.

Input and Output:

- ➢ 1 text prompt + N camera parameters
- ➢ N synchronized videos.

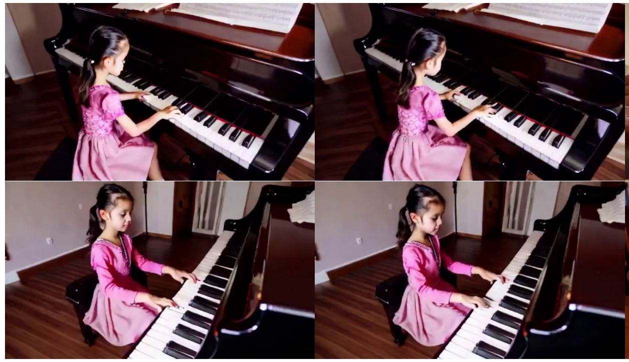
Main Features:

- ✓ Multi-camera synchronized video generation.
- ✓ Enable synthesis from diverse viewpoints.
- ✓ A simple and efficient module on top of pre-trained text-to-video models.

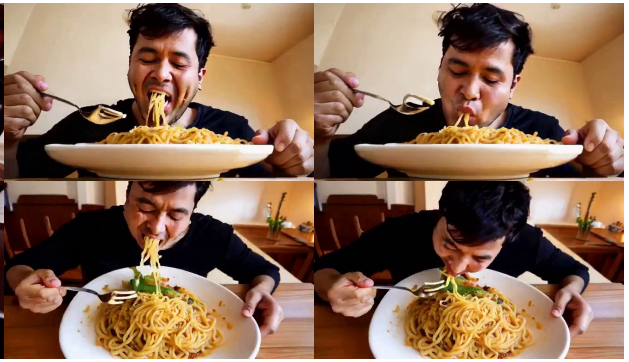
SynCamVideo Dataset:

Release a multi-camera synchronized video dataset rendered with Unreal Engine 5.

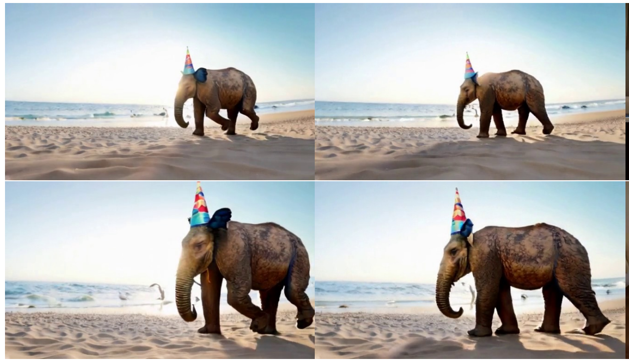
90° Difference in Azimuth



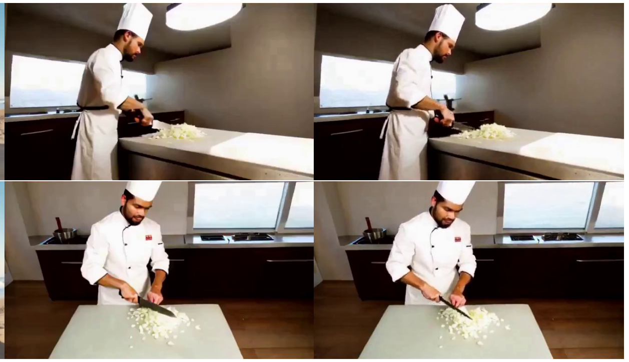
45° Difference in Elevation



Close-Up and Wide Shot



60° in Azimuth+30° in Elevation

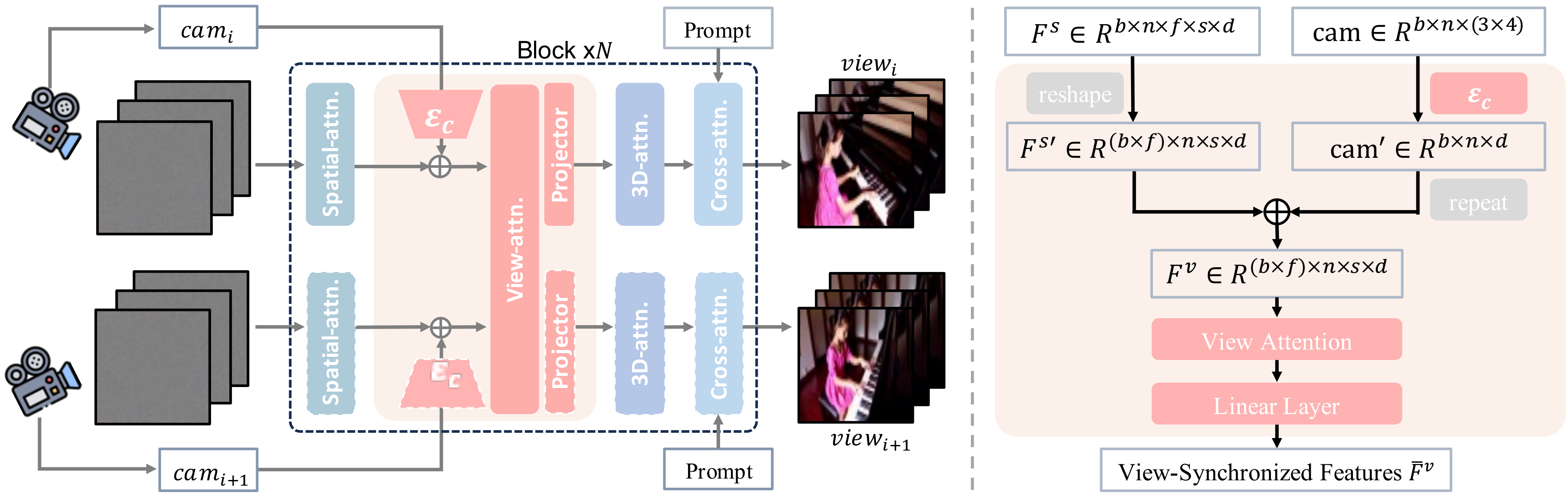


Background & Motivation

- Multi-View Video Generation.
  - Existing works primarily focus on 4D object generation or generation on a specific domain (e.g., autonomous driving).
  - This paper explores how to achieve open-domain multi-camera video generation.

- Why Multi-View Open-Domain Generation?
  - In filmmaking, switching back and forth between multiple cameras is commonly used to create a storytelling atmosphere.
  - It can be used as a data generator for various downstream tasks (e.g., robotic manipulation, 3D human pose estimation).

Method



cam_i
Prompt
Block xN
$view_i$
Spatial-attn.
$\mathcal{E}_c$
View-attn.
Projector
3D-attn.
Cross-attn.
Spatial-attn.
$\mathcal{E}_c$
Projector
3D-attn.
Cross-attn.
$view_{i+1}$
cam_{i+1}
Prompt
$F^s \in R^{b \times n \times f \times s \times d}$
$\text{cam} \in R^{b \times n \times (3 \times 4)}$
reshape
$\mathcal{E}_c$
$F^{s\prime} \in R^{(b \times f) \times n \times s \times d}$
$\text{cam}' \in R^{b \times n \times d}$
repeat
$F^v \in R^{(b \times f) \times n \times s \times d}$
View Attention
Linear Layer
View-Synchronized Features $\bar{F}^v$

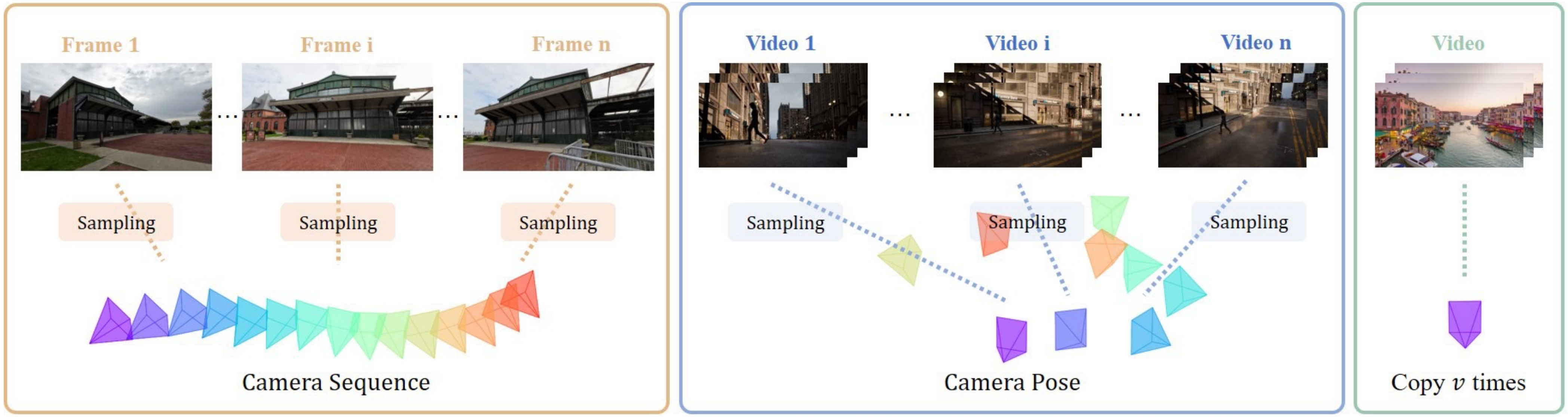
(a) Overview of SynCamMaster **(b) Multi-View Synchronization Module**

- Based on the pre-trained text-to-video model, two components are newly introduced:
  - The camera encoder projects the camera extrinsics into the embedding space.
  - The multi-view synchronization module, as plugged in each TransformerBlock, modulates multi-view features under the guidance of camera parameters.

$$\mathbf{F}_i^v = \mathbf{F}_i^s + \mathcal{E}_c(\mathtt{cam}^i), \tag{5}$$

$$\overline{\mathbf{F}}_i^v = \mathbf{F}_i^v + \mathtt{projector}(\mathtt{attn\_view}(\mathbf{F}_1^v, \ldots, \mathbf{F}_n^v)[i]), \tag{6}$$

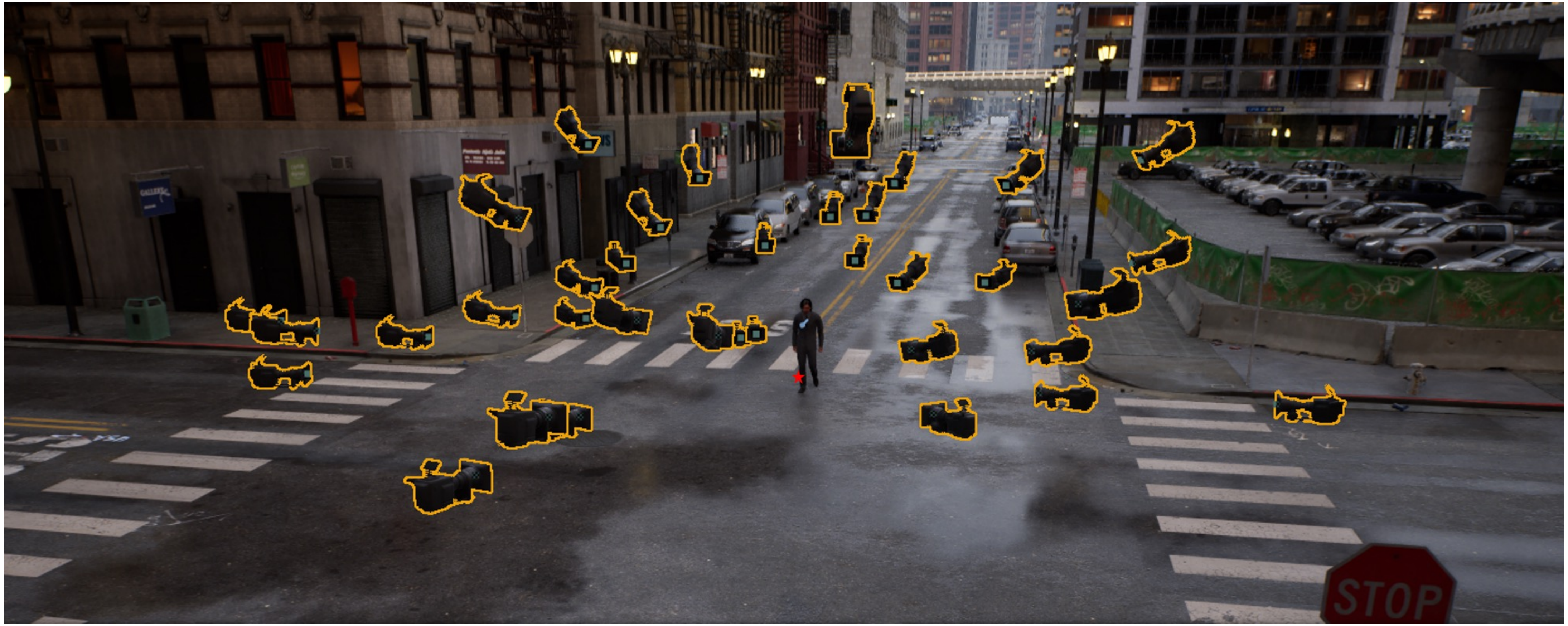
Training Data



(a). Construction of Multi-view Image Data (b). Rendered Multi-view Video Data (c). General Videos

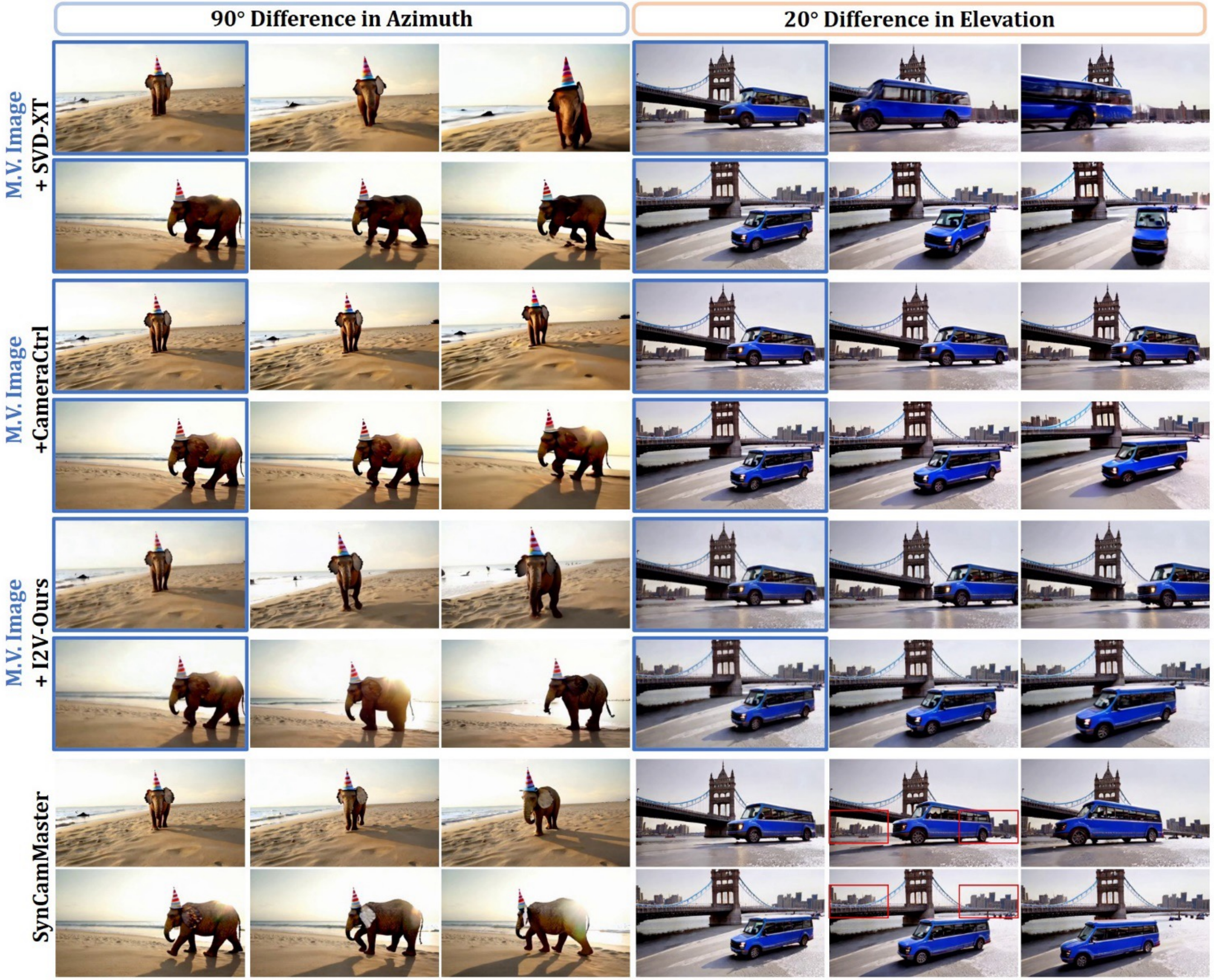
- Due to the scarcity of available multi-view videos, we used a hybrid training set to enhance the model's robustness and improve the visual quality of the generated videos, the training set is composed of:
  - Multi-view image data from videos with camera movements.
  - Multi-view video data from the rendered SynCamVideo Dataset.
  - General video data from the internet.

SynCamVideo Dataset



- Multi-Camera Synchronized Videos + Corresponding Camera Parameters
- Rendered with UnrealEngine 5

Results



An elephant wearing a colorful birthday hat is walking along the sandy beach.

A blue bus drives across the iconic Tower Bridge in London.

Results

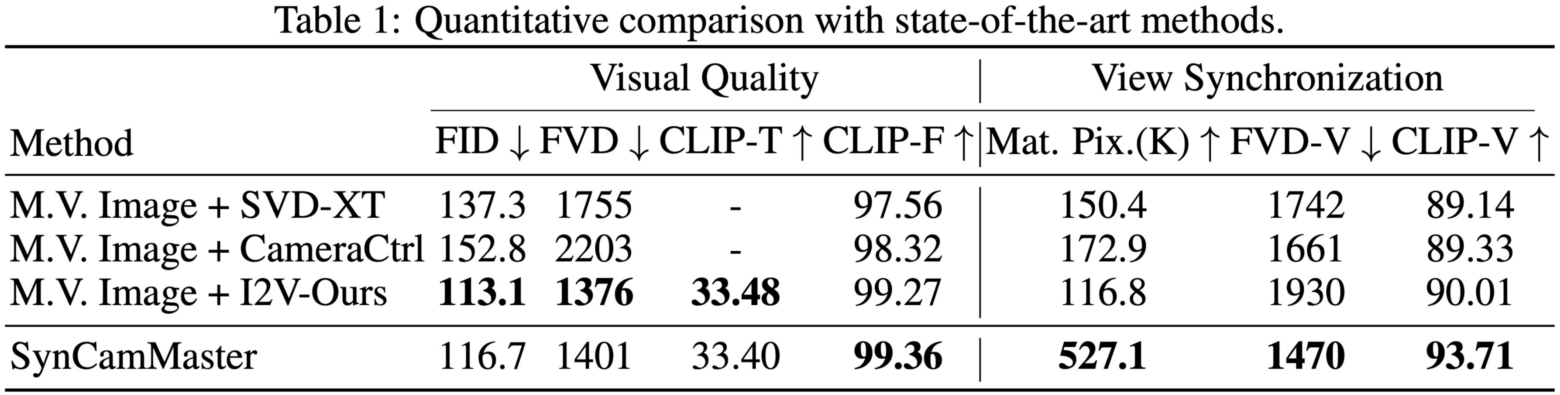


Table 1: Quantitative comparison with state-of-the-art methods.

	Visual Quality				View Synchronization		
Method	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
M.V. Image + SVD-XT	137.3	1755	-	97.56	150.4	1742	89.14
M.V. Image + CameraCtrl	152.8	2203	-	98.32	172.9	1661	89.33
M.V. Image + I2V-Ours	**113.1**	**1376**	**33.48**	99.27	116.8	1930	90.01
SynCamMaster	116.7	1401	33.40	**99.36**	**527.1**	**1470**	**93.71**

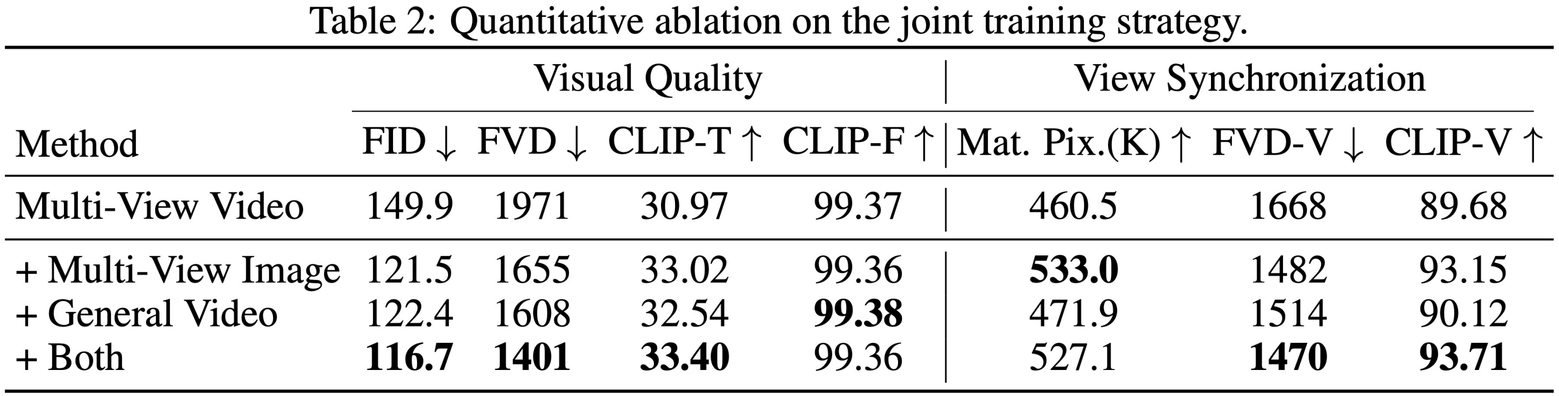


Table 2: Quantitative ablation on the joint training strategy.

	Visual Quality				View Synchronization		
Method	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
Multi-View Video	149.9	1971	30.97	99.37	460.5	1668	89.68
+ Multi-View Image	121.5	1655	33.02	99.36	**533.0**	1482	93.15
+ General Video	122.4	1608	32.54	**99.38**	471.9	1514	90.12
+ Both	**116.7**	**1401**	**33.40**	99.36	527.1	**1470**	**93.71**

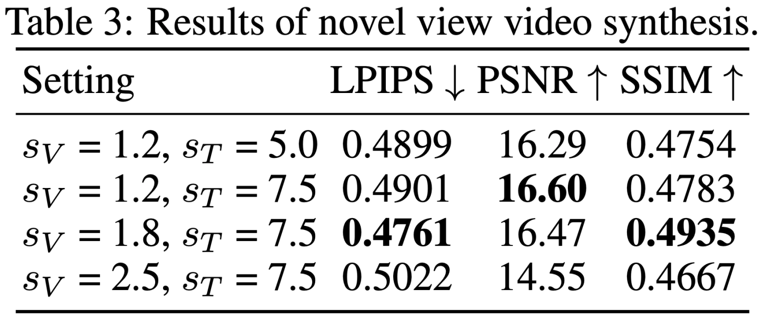


Table 3: Results of novel view video synthesis.

Setting	LPIPS ↓	PSNR ↑	SSIM ↑
$s_V = 1.2, s_T = 5.0$	0.4899	16.29	0.4754
$s_V = 1.2, s_T = 7.5$	0.4901	**16.60**	0.4783
$s_V = 1.8, s_T = 7.5$	**0.4761**	16.47	**0.4935**
$s_V = 2.5, s_T = 7.5$	0.5022	14.55	0.4667

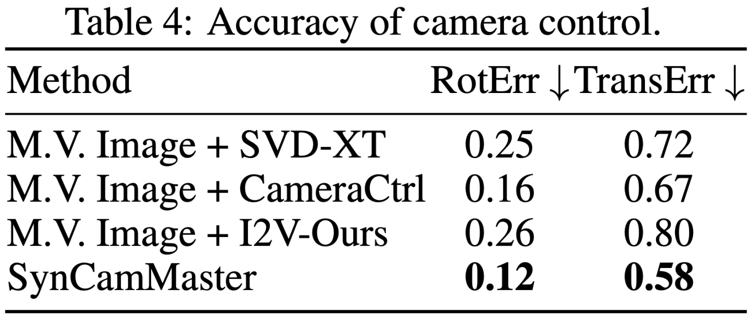
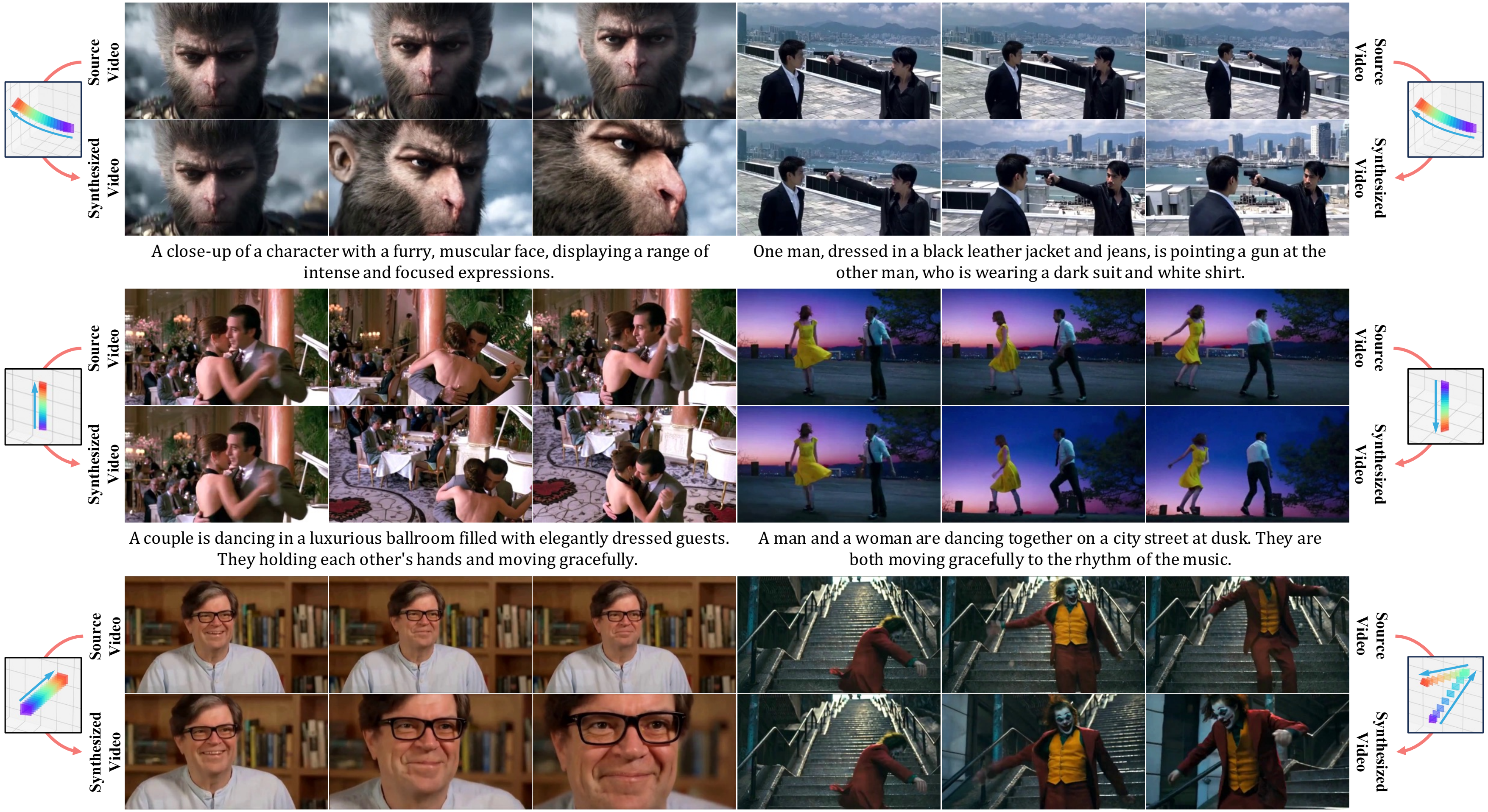


Table 4: Accuracy of camera control.

Method	RotErr ↓	TransErr ↓
M.V. Image + SVD-XT	0.25	0.72
M.V. Image + CameraCtrl	0.16	0.67
M.V. Image + I2V-Ours	0.26	0.80
SynCamMaster	**0.12**	**0.58**

Subsequent Work: ReCamMaster



A close-up of a character with a furry, muscular face, displaying a range of intense and focused expressions.

One man, dressed in a black leather jacket and jeans, is pointing a gun at the other man, who is wearing a dark suit and white shirt.

A couple is dancing in a luxurious ballroom filled with elegantly dressed guests. They holding each other's hands and moving gracefully.

A man and a woman are dancing together on a city street at dusk. They are both moving gracefully to the rhythm of the music.

A middle-aged man wearing a light blue shirt. He is smiling warmly and appears to be in a relaxed and happy mood.

A person in a vibrant red suit, bright red clown wig, and white face paint is descending a flight of stairs.

- Input: source video + target camera trajectory.
- Output: Video with the novel camera trajectory.

Take Home Messages

- We propose SynCamMaster to synthesize multi-camera videos from the text prompt and camera extrinsic.
- We release a multi-camera synchronized video dataset rendered with Unreal Engine 5.
- Our subsequent work, ReCamMaster, can recapture an input video using novel camera trajectories.
- For more information:

SynCamMaster Project Page

ReCamMaster Project Page

Thanks for your attention!