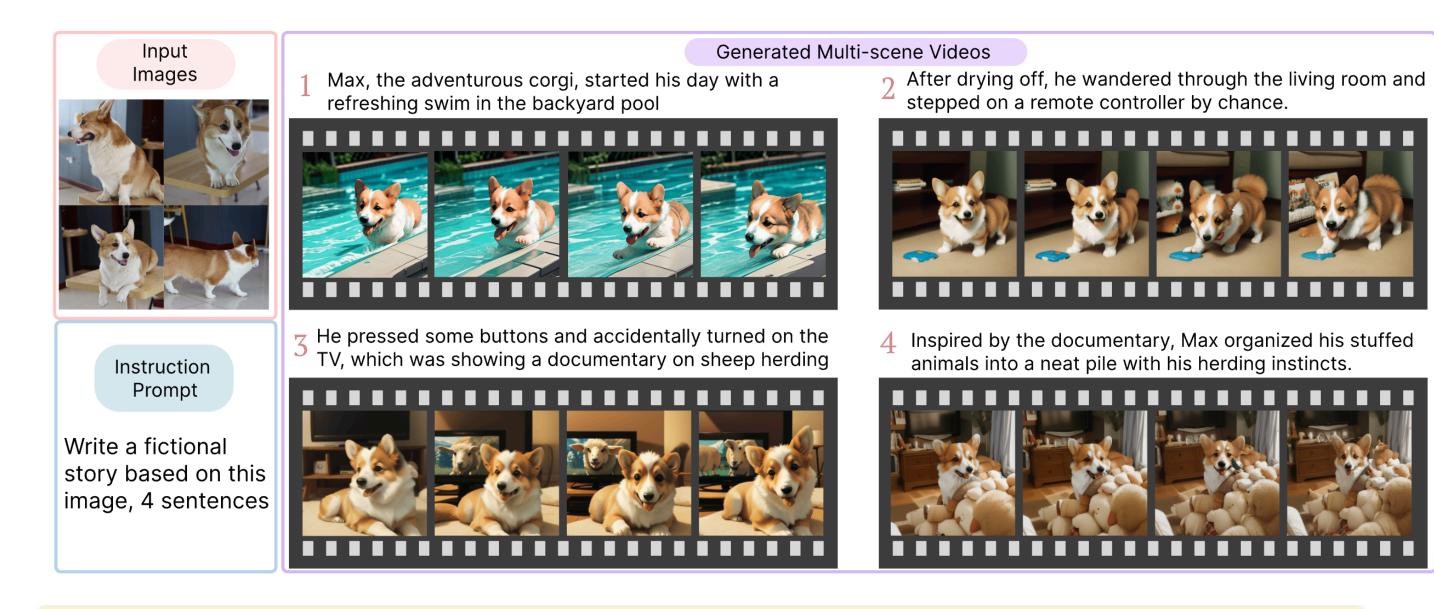


Corgi: Cached Memory Guided Video Generation

Xindi Wu^{1, 2}, Uriel Singer¹, Zhaojiang Lin¹, Andrea Madotto¹, Xide Xia¹ Yifan Ethan Xu¹, Paul A. Crook¹, Xin Luna Dong¹, Seungwhan Moon¹ ¹FAIR, Meta & Meta Reality Labs ²Princeton University X @cindy x wu xindiw@princeton.edu



Multi-Scene Video Generation



Multi-scene video generation, the process of generating multi-scene long videos with multimodal inputs, primarily faces challenges in consistency, faithfulness, and diversity.

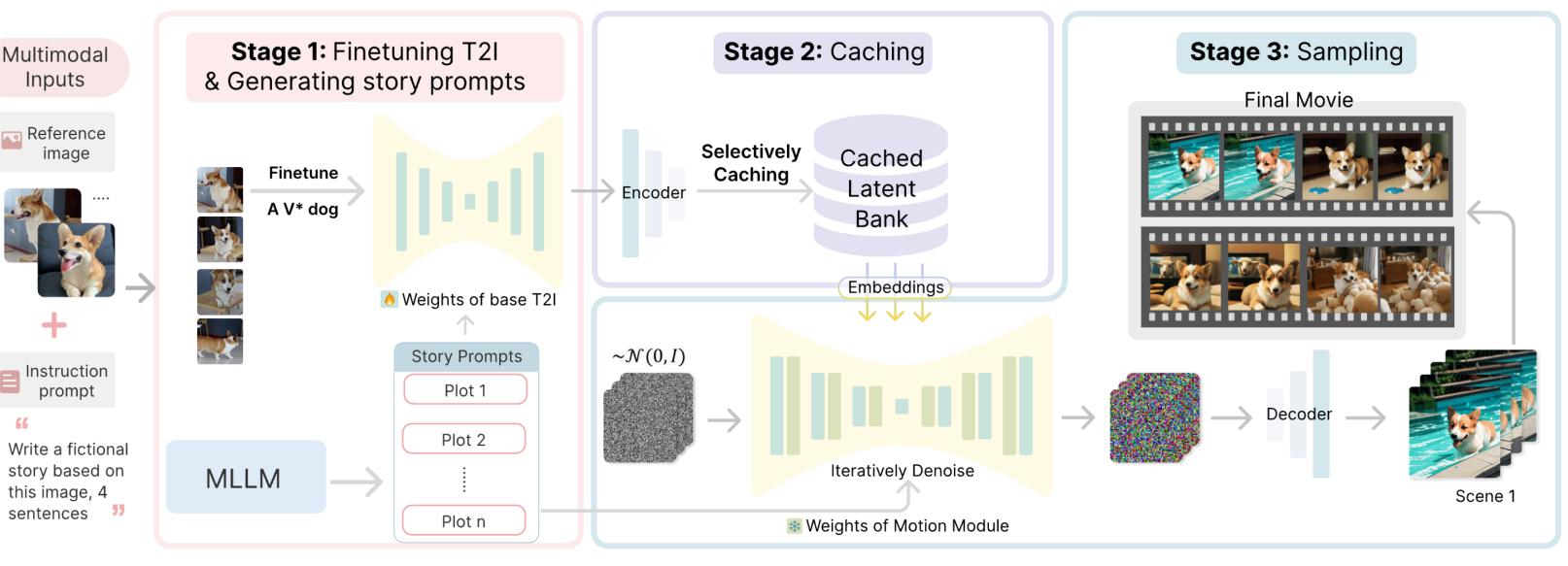
- Core insights: Repeated key moments trigger similar brain activations, helping viewers grasp the storyline.
- We propose a multi-scene video generation method that generates key frames first, treating them as core memories stored in a cached latent memory bank.

How can we create multi-scene videos that are consistent, faithful, and diverse?

Corgi

Stage 1 Finetuning T2I & Generating Story Prompts

- A Multimodal-LLM (MLLM) generates story prompts from 3-5 reference images and an instruction prompt.
- The fine-tuned T2I model creates intermediate images based on these prompts.



Stage 2 Caching

- Latents from a pre-trained encoder are stored in a cached memory, serving as the basis of the initial image conditioning and, along with the story prompts, guide the video generation process.
- **Coverage Caching** in the VAE latent space maximizes latent variety while staying compact and flexible to avoid repetitiveness and improve diversity in backgrounds, poses, and more.

Coverage Score: $D = \|\mathbf{z}_{new} - \mathbf{z}_{centroid}\|$

 $\mathbf{z}_{\text{centroid}} = \frac{1}{r} \sum_{i=1}^{r} \mathbf{z}_i$ is the center of all existing cached latents.

Stage 3 Sampling

• Motion dynamics are added using a temporal transformer, producing the final multi-scene video by stitching together generated clips.

Cached Latent Conditioning: To condition on the cached latent signals during the video generation process, we add weighted \mathbf{Z}_i

$$\hat{\boldsymbol{\epsilon}} = \{\epsilon_1 + \lambda_1 \mathbf{z}_i, \epsilon_2 + \lambda_2 \mathbf{z}_i, ..., \epsilon_N + \lambda_N \mathbf{z}_i\}$$

 $\{\lambda_k\}_{k=1}^N$:weights that control how much influence the cached latent have on the generation of subsequent frames.

Results

Generated Results

• Baseline comparisons

Method	Consistency (\downarrow)		Faithfulness (\uparrow)		Diversity (\uparrow)	
	Short-term	Long-term	Visual	Textual		-
Gen-L-Video [8]	30.53 ± 7.41	28.51 ± 5.49	_	32.76 ± 3.49	42.26 ± 2.98	
FreeNoise [6]	28.97 ± 4.12	32.83 ± 7.33	—		49.12 ± 5.92	
Corgi (ours)	$\textbf{12.58} \pm 5.76$	$\textbf{11.63} \pm \textbf{5.23}$	$\textbf{85.83} \pm 6.38$	$\textbf{37.11} \pm \textbf{4.27}$	$\textbf{52.84} \pm \textbf{3.28}$	

Ablation Cached Latent Selection

	Consistency (↓)		Faithfulness (†)		Diversity (†)
Cached Latents	Short-term	Long-term	Visual	Textual	
Random	11.64 ± 5.89	$\textbf{10.85} \pm 6.71$	85.33 ± 5.91	$\textbf{36.58} \pm \textbf{3.49}$	40.27 ± 4.12
Selected	12.58 ± 5.76	11.63 ± 5.23	$\textbf{85.83} \pm 6.38$	$\textbf{37.11} \pm 4.27$	$\textbf{52.84} \pm 3.28$

<u>Cached Latent Conditioning</u>

	Consistency (\downarrow)		Faithfulness ([†])		Diversity (†)
Weight Setting	Short-term	Long-term	Visual	Textual	
Constant	$\textbf{7.42} \pm 4.37$	17.93 ± 5.02	$\textbf{86.44} \pm 8.24$	$\textbf{35.94} \pm 5.73$	$\textbf{38.64} \pm \textbf{6.74}$
Low	21.36 ± 6.15	$\textbf{23.48} \pm \textbf{4.63}$	$\textbf{75.89} \pm \textbf{8.06}$	$\textbf{32.18} \pm \textbf{7.93}$	49.27 ± 5.15
High	$\textbf{8.57} \pm 5.82$	$\textbf{25.14} \pm \textbf{4.85}$	$\textbf{54.38} \pm \textbf{9.53}$	$\textbf{21.49} \pm \textbf{3.81}$	$\textbf{34.96} \pm \textbf{7.36}$
Linear (ours)	12.58 ± 5.76	$\textbf{11.63} \pm 5.23$	$\textbf{85.83} \pm \textbf{6.38}$	$\textbf{37.11} \pm 4.27$	$\textbf{52.84} \pm 3.28$

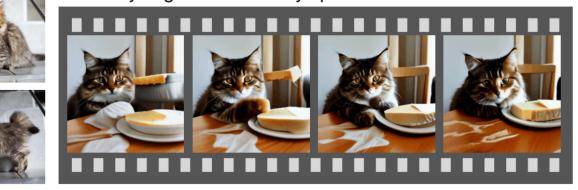
Input Images

Generated Multi-scene videos

1 Whiskers the cat found her bowl empty, her stomach grumbling a little tune of hunger



³ Her eyes gleamed as they spotted a cheese



1 A teddy bear, still sleepy-eyed, wakes up in his cozy little bed, stretching out his fluffy arms



³ At the mall, he wanders around a bit, checking out the colorful store windows



2 She prowled the kitchen, sniffing the air for a stray crumb or a forgotten treat



4 With a swift paw, she fished it out, and the cheese was a satisfying memory



2 He hops onto his red bike, pedaling through the bustling streets to see his friends



4 There, he waits for his friends' arrival at the window, wondering what adventures they'll have today

