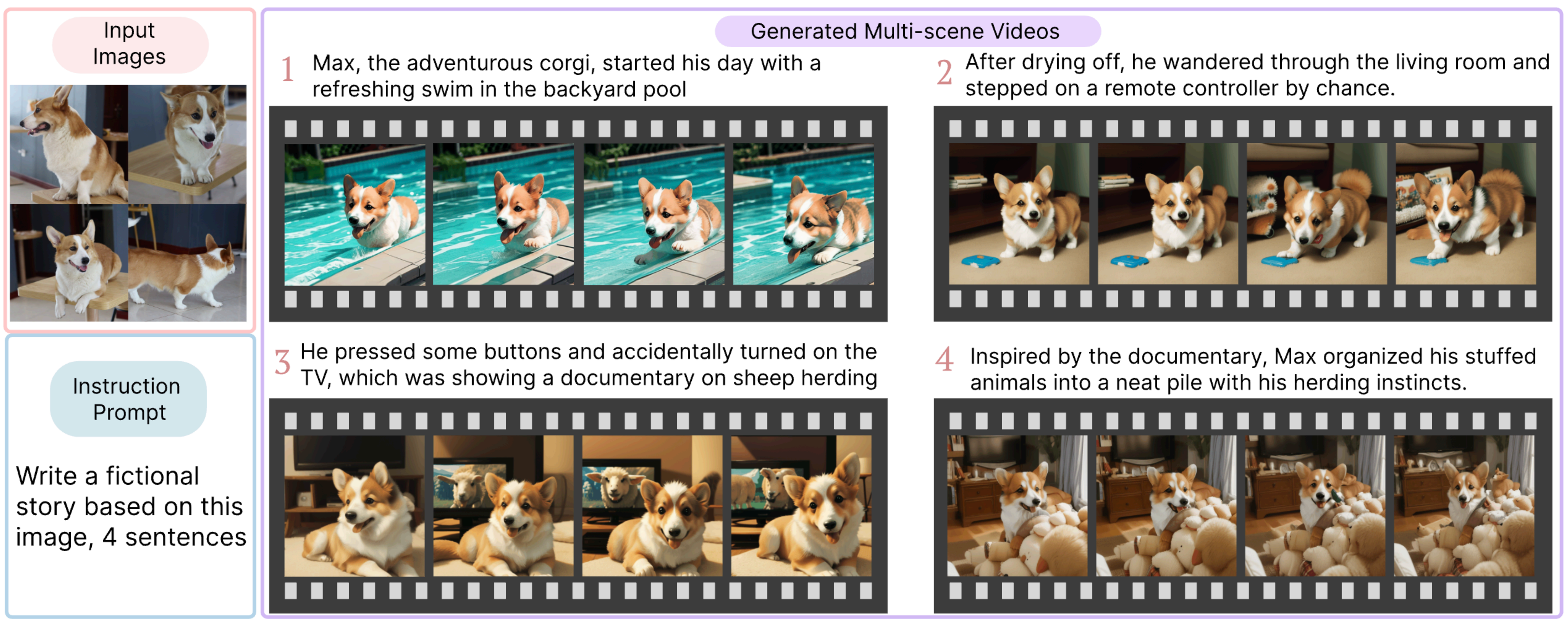
Corgi: Cached Memory Guided Video Generation

Xindi Wu[1,2], Uriel Singer[1], Zhaojiang Lin[1], Andrea Madotto[1], Xide Xia[1]
Yifan Ethan Xu[1], Paul A. Crook[1], Xin Luna Dong[1], Seungwhan Moon[1]
[1]FAIR, Meta & Meta Reality Labs [2]Princeton University
xindiw@princeton.edu @cindy_x_wu

Paper

Multi-Scene Video Generation



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

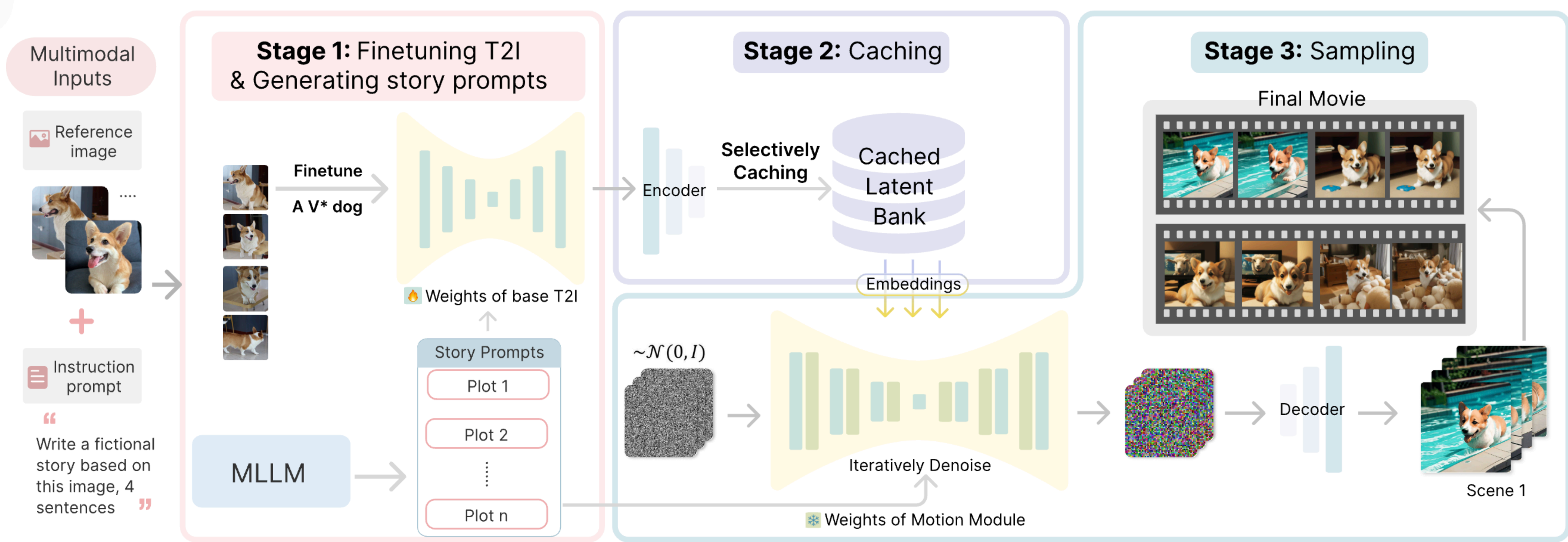
*Multi-scene video generation, the process of generating multi-scene long videos with multimodal inputs, primarily faces challenges in **consistency, faithfulness, and diversity.***

- 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.

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

$\mathbf{z}_{\text{centroid}} = \frac{1}{r}\sum_{i=1}^{r}\mathbf{z}_i$ 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 $\mathbf{z}_i$

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

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

Results

- Baseline comparisons

Method	Consistency (↓) Short-term	Consistency (↓) Long-term	Faithfulness (↑) Visual	Faithfulness (↑) Textual	Diversity (↑)
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	–	21.18 ± 0.48	49.12 ± 5.92
Corgi (ours)	**12.58** ± 5.76	**11.63** ± 5.23	**85.83** ± 6.38	**37.11** ± 4.27	**52.84** ± 3.28

- Ablation

Cached Latent Selection

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

Cached Latent Conditioning

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

Generated Results