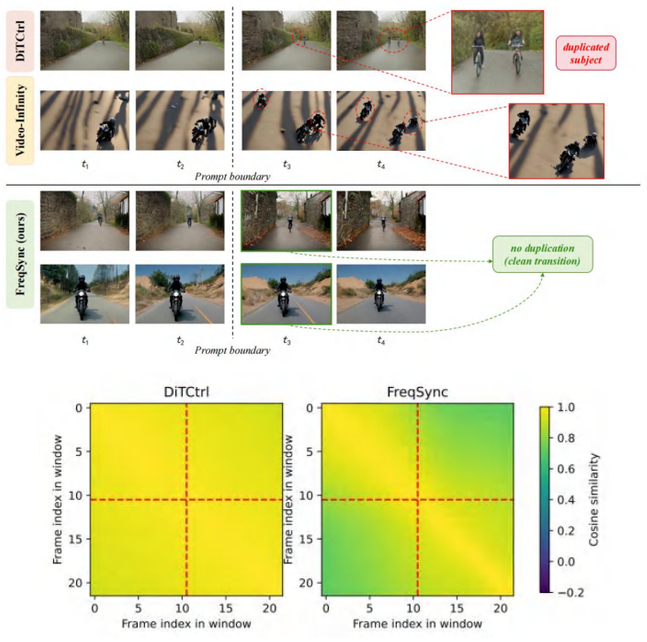
Frequency-Synchronized Boundary Coupling for Training-Free Multi-Prompt Long Video Generation

Yipan Xu[1], Shuyong Gao[2]*, Jiyuan Fu[1], Keliang Yin[1], Jiayu Chen[1], Qianyu Guo[3], Yang Liu[4], Wenqiang Zhang[1]*
[1]Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
[2]Naval Medical Center of PLA, Second Military Medical University [3]Shanghai Institute of Virology, Shanghai Jiao Tong University
[4]College of Electronic and Information Engineering, Tongji University * Corresponding author.

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

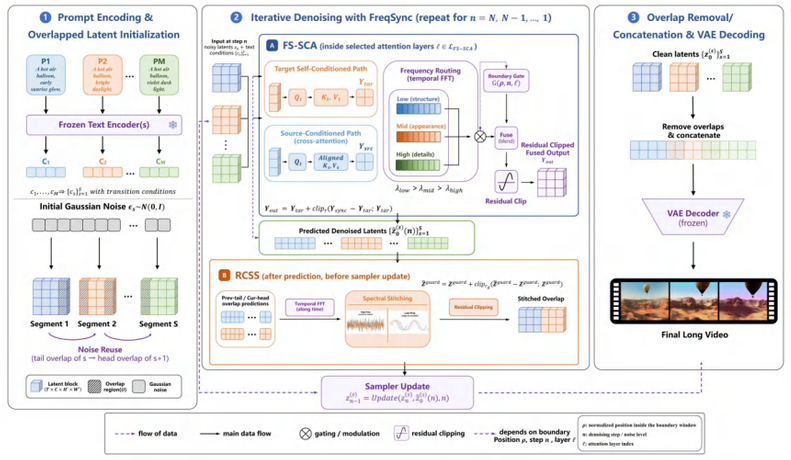


Problem. Weak coupling causes flicker, background jumps, and discontinuous motion, while overly strong coupling may over-preserve source-side layouts and semantics. Such source retention can produce ghosting or duplicated subjects near boundaries.

Observation. FreqSync preserves local diagonal continuity while reducing excessive cross-boundary similarity, supporting frequency-selective boundary coupling between adjacent video segments.

Key idea. Source-side low- and mid-frequency components should provide layout, illumination, and motion continuity, while target-side high-frequency components should preserve the new visual details introduced by the target prompt.

Method



We propose FreqSync, a training-free framework for multi-prompt long video generation. It first encodes prompt conditions and initializes neighboring segments with overlap noise reuse. During denoising, FS-SCA operates inside selected attention layers, while RCSS stitches denoised overlap predictions before the sampler update. After denoising, overlapped regions are removed, the cleaned latent segments are concatenated, and the frozen VAE decoder produces the final long video.

Frequency-Synchronized Source-Conditioned Attention (FS-SCA)

FS-SCA performs frequency-selective source conditioning inside selected attention layers, enabling the target segment to inherit boundary-relevant structure without copying full-frequency source features.

Residual-Clipped Spectral Stitching (RCSS)

RCSS performs spectral seam alignment and conservative low-frequency continuation on denoised overlap predictions under residual constraints.

Experiments

Quantitative comparison

Method	Seed	CSCV (%)↑	Motion Smooth. (%)↑	Text-Image Similarity↑	Visual Quality↑	T2V Align.↑	Physical Consist.↑
MEVG [33]	42	68.09	95.54	29.86	1.957	2.778	2.261
MEVG [33]	123	67.23	96.04	29.29	2.008	2.868	2.306
MEVG [33]	999	66.34	95.39	29.52	1.946	2.674	2.272
MEVG [33]	Average	67.22	95.66	29.56	1.970	2.773	2.280
Video-Infinity [42]	42	83.72	96.53	29.92	3.255	3.511	3.200
Video-Infinity [42]	123	83.63	96.38	29.75	3.210	3.347	3.162
Video-Infinity [42]	999	83.52	96.42	29.49	3.275	3.409	3.268
Video-Infinity [42]	Average	83.62	96.45	29.72	**3.247**	3.422	3.210
DiTCtrl [7]	42	86.46	98.48	30.35	3.129	3.753	3.230
DiTCtrl [7]	123	87.45	98.59	30.48	3.171	3.801	3.275
DiTCtrl [7]	999	84.42	98.25	30.14	3.142	3.721	3.181
DiTCtrl [7]	Average	86.11	**98.44**	30.33	3.147	3.758	3.228
FreqSync (Ours)	42	87.39	98.19	30.50	3.244	3.864	3.327
FreqSync (Ours)	123	87.00	98.42	30.58	3.203	3.854	3.191
FreqSync (Ours)	999	84.91	98.15	30.51	3.195	3.829	3.203
FreqSync (Ours)	Average	**86.43**	98.25	**30.53**	3.214	**3.849**	**3.240**

Cross-backbone evaluation

Backbone	Method	CSCV (%)	Motion (%)	Text–Image	VQ	T2V	Physical
VideoCrafter2	DiTCtrl	81.04	97.15	34.01	**3.095**	**3.665**	**3.057**
VideoCrafter2	FreqSync	**86.50**	**99.35**	**34.23**	2.836	3.576	2.842
Wan2.1-1.3B	DiTCtrl	84.05	**98.25**	31.91	3.267	3.900	3.404
Wan2.1-1.3B	FreqSync	**87.00**	98.04	**33.50**	**3.872**	**4.095**	**3.809**

Ablation studies

Variant	FS-SCA	RCSS	CSCV (%)↑	Text-Image Sim.↑
Isolated	×	×	76.96	**30.89**
w/o FS-SCA (RCSS only)	×	✓	85.73	30.62
w/o RCSS (FS-SCA only)	✓	×	82.08	30.56
Full FreqSync	✓	✓	**89.14**	30.71

Frequency-routing ablation

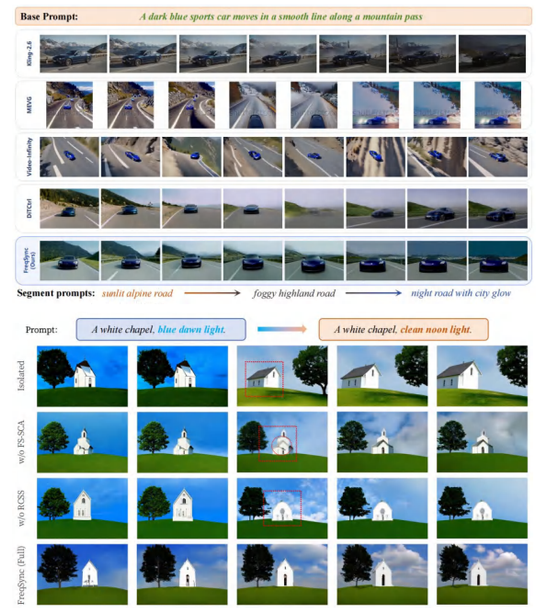
Variant	Source routing (L, M, H)	CSCV (%)	Text–Image Similarity	T2V Alignment
Full FreqSync	(0.98, 0.68, 0.035)	$\mathbf{89.139 \pm 6.774}$	34.013 ± 3.019	$\mathbf{3.951 \pm 0.370}$
Low+Mid-only	(0.98, 0.68, 0)	88.956 ± 7.121	$\mathbf{34.027 \pm 3.028}$	3.891 ± 0.369
High-only	(0, 0, 0.98)	86.302 ± 7.335	33.913 ± 3.064	3.805 ± 0.418
Reverse Routing	(0.035, 0.68, 0.98)	85.795 ± 8.250	33.885 ± 2.939	3.821 ± 0.349
Uniform Routing	(0.565, 0.565, 0.565)	88.380 ± 7.506	33.964 ± 3.075	3.903 ± 0.411
w/o FS-SCA	—	85.693 ± 8.385	33.570 ± 2.902	3.905 ± 0.361

Human evaluation

Method	Target semantic correctness↑	Source leakage and artifacts↑	Transition smoothness↑	Visual quality↑	Overall preference↑
MEVG [33]	1.67	1.53	1.70	1.83	1.63
Video-Infinity [42]	2.70	2.57	2.83	2.97	2.70
DiTCtrl [7]	3.33	3.07	2.80	2.90	3.17
FreqSync (Ours)	**4.30**	**4.23**	**4.27**	**4.07**	**4.30**

Quantitative evaluation ranks FreqSync first on four of six metrics. **Ablations** show complementary roles for FS-SCA and RCSS, with the full model achieving the best CSCV. **Cross-backbone results** support transfer of the boundary-coupling principle with architecture-dependent trade-offs. **Human evaluation** yields the highest mean scores across all five dimensions.

Qualitative results

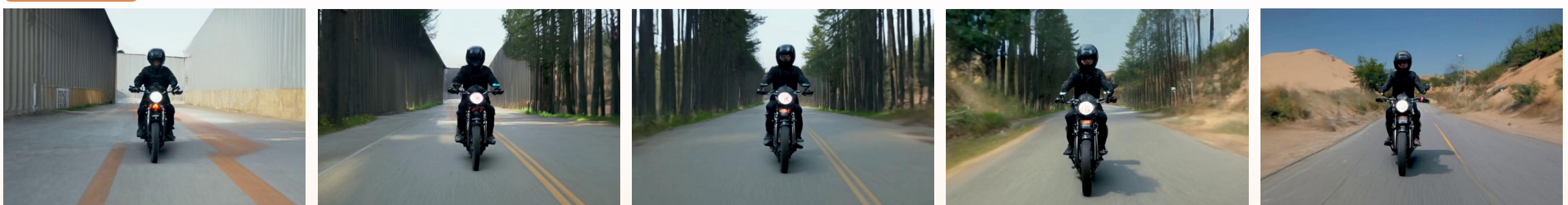


Prompt 1 *A small rowboat drifts gently across calm water*

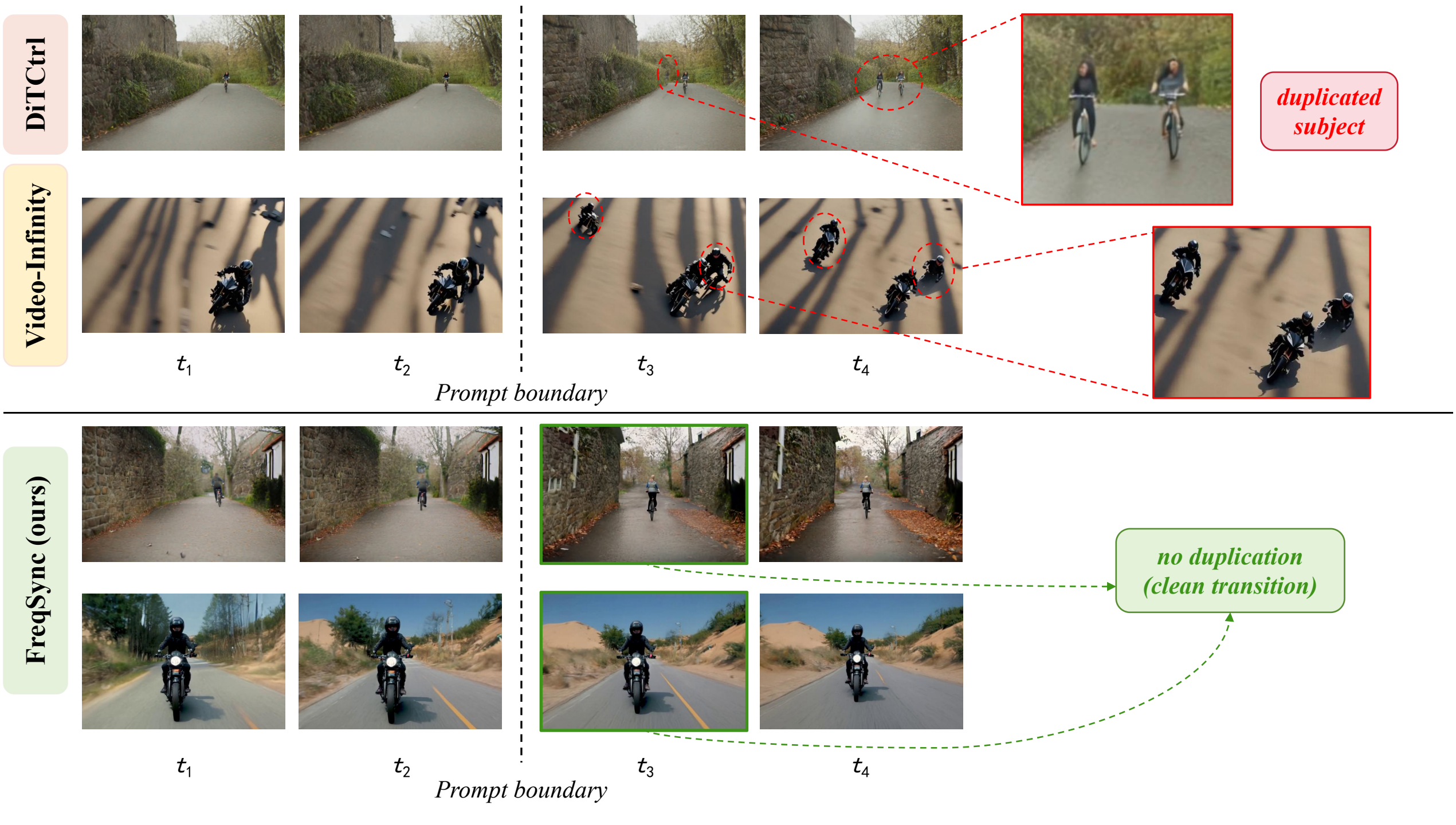


reed-lined marsh → lotus pond → mirror-like salt flat pool

Prompt 2 *A motorcyclist in black gear rides forward down a straight route*



industrial district road → forest highway → desert highway


DiTCtrl
Video-Infinity
duplicated subject
t_1
t_2
t_3
t_4
Prompt boundary
FreqSync (ours)
no duplication (clean transition)
t_1
t_2
t_3
t_4
Prompt boundary

Base Prompt: *A dark blue sports car moves in a smooth line along a mountain pass*

Kling-2.6



MEVG



Video-Infinity



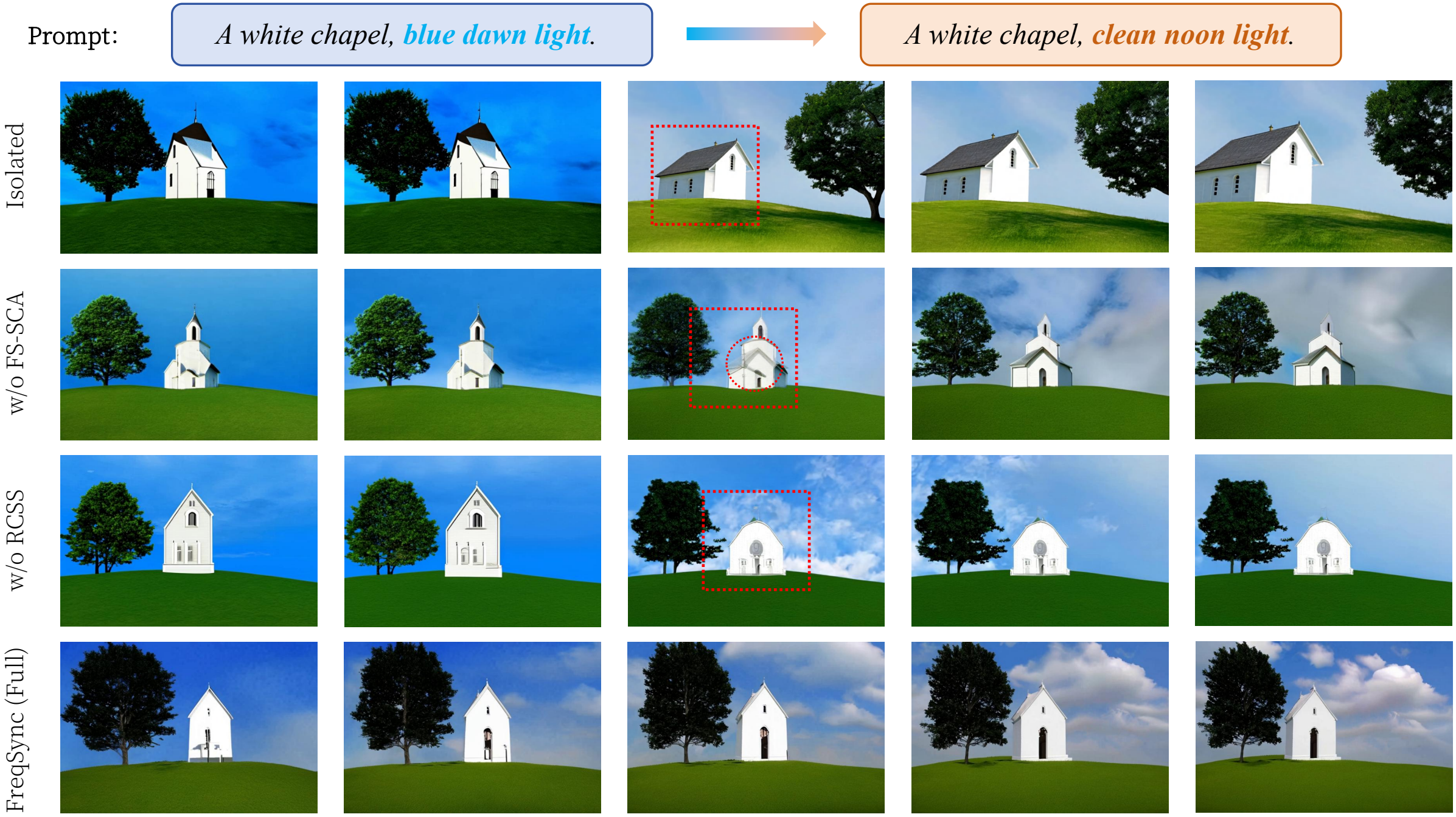
DiTCtrl



FreqSync (Ours)



Segment prompts: *sunlit alpine road* → *foggy highland road* → *night road with city glow*

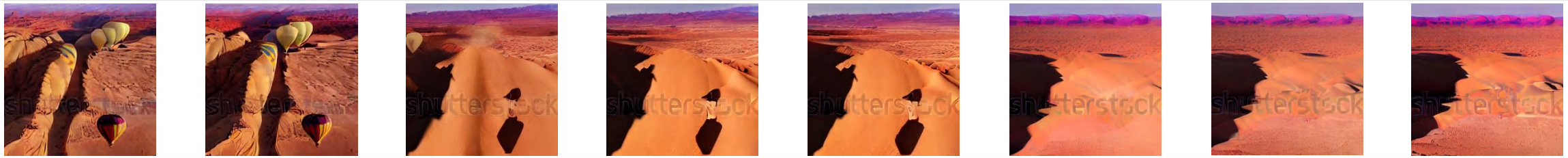

Prompt:
A white chapel, blue dawn light.
A white chapel, clean noon light.
Isolated
w/o FS-SCA
w/o RCSS
FreqSync (Full)

Base Prompt: *A hot air balloon*

Kling-2.6



MEVG



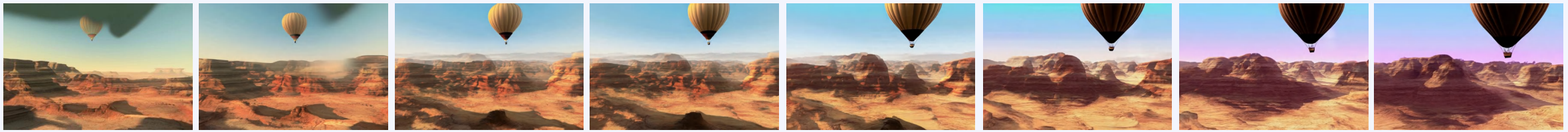
Video-Infinity



DiTCtrl



FreqSync (Ours)



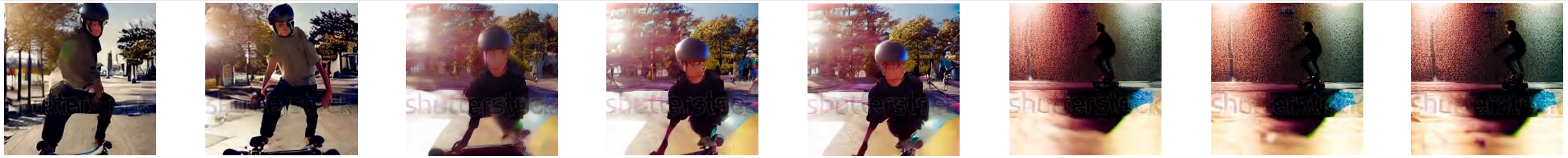
Segment prompts: *early sunrise glow* → *bright daylight* → *violet dusk light*

Base Prompt: *A young skateboarder glides forward through an open public square*

Kling-2.6



MEVG



Video-Infinity



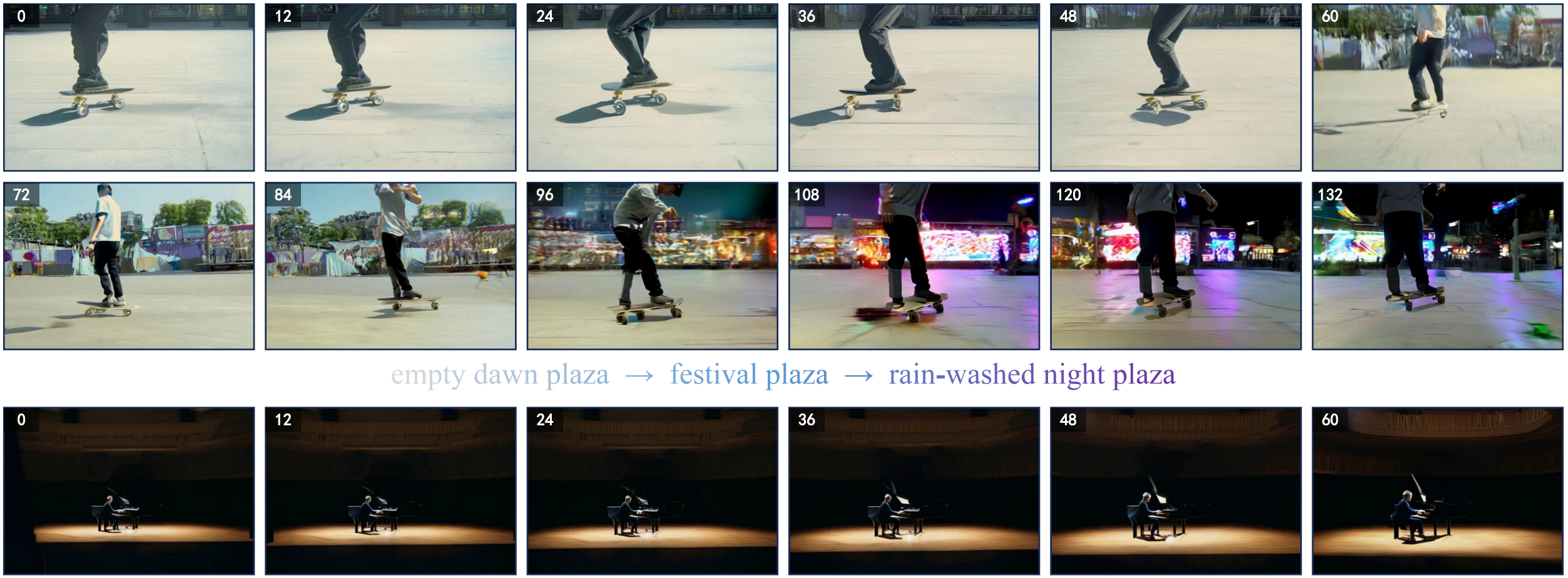
DiTCtrl



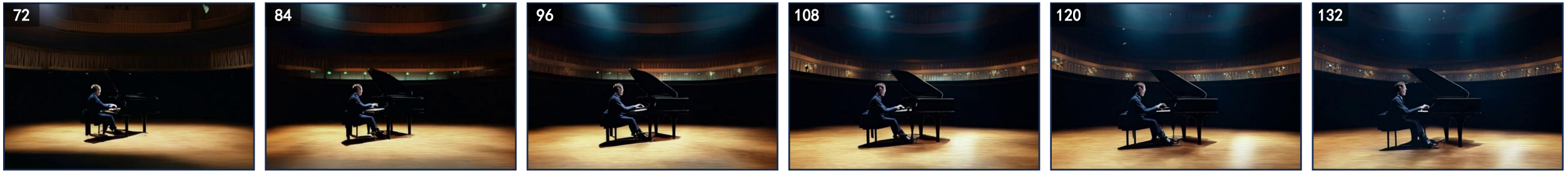
FreqSync (Ours)



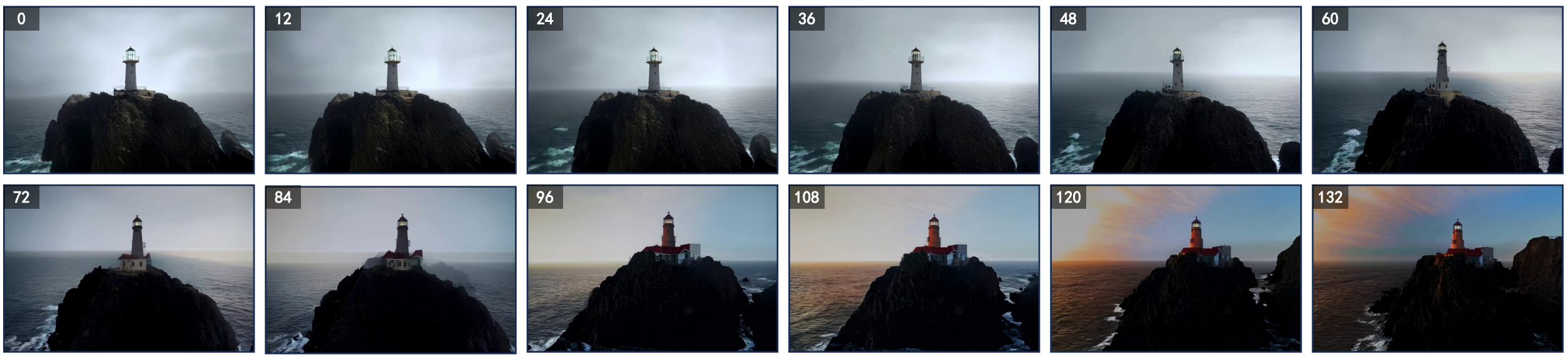
Segment prompts: *empty dawn plaza* → *festival plaza* → *rain-washed night plaza*



empty dawn plaza → festival plaza → rain-washed night plaza



dim rehearsal worklight → balanced concert spotlight → glittering gala lighting



gray storm light → silver post-rain light → orange sunset light