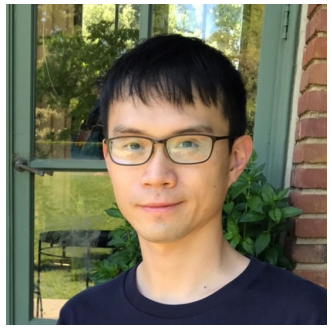
IllumiCraft: Unified Geometry and Illumination Diffusion for Controllable Video Generation


Yuanze Lin[1]


Yi-Wen Chen[3]


Yi-Hsuan Tsai[4]


Ronald Clark[1]


Ming-Hsuan Yang[2,5]

1 - University of Oxford
2 - UC Merced
3 - NEC Labs America
4 - Atmanity Inc.
5 - Google DeepMind

Motivation

- Previous video relighting approaches:

  (1) Explicitly omit geometric guidance, leading to the loss of lighting fidelity and temporal coherence when the scene's geometry changes.

  (2) Haven't presented a comprehensive video dataset with illumination and 3D geometry cues.

- **Our Method:**

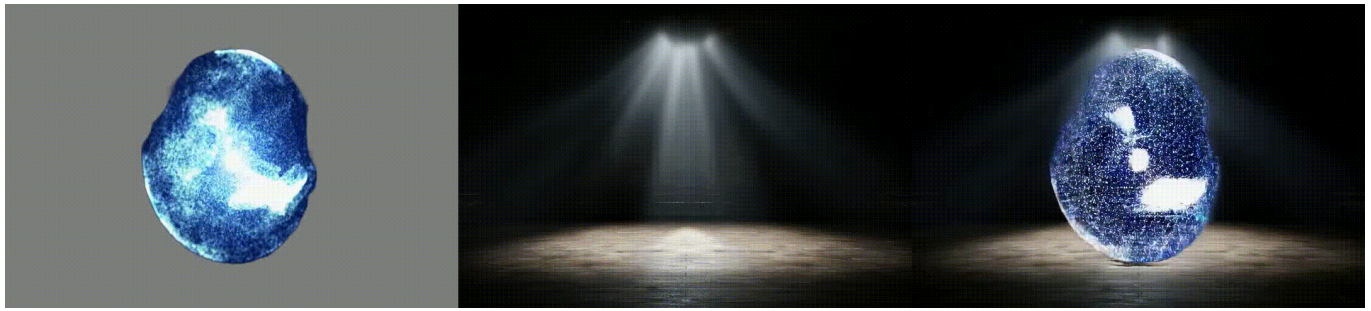
  (1) Unifies illumination and geometry guidance for high-quality video relighting. It supports text-conditioned and background-conditioned relighting for videos.

  (2) Introduces a high-quality video dataset comprising 20,170 video pairs to support video relighting and serves as a valuable resource for broader controllable video generation tasks.

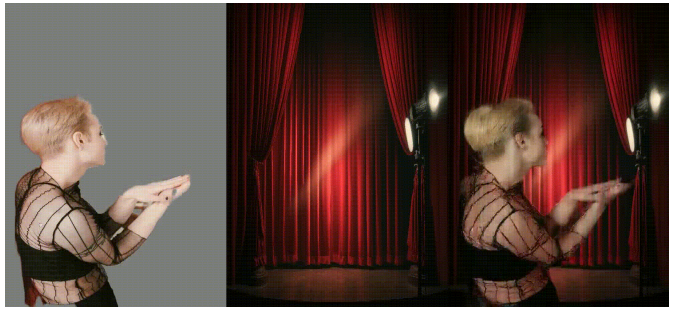
  (3) Has been demonstrated effectiveness against state-of-the-art methods on the video relighting task.

Overview

"A glowing, translucent sphere, ···, crisp radiant overhead, ···"

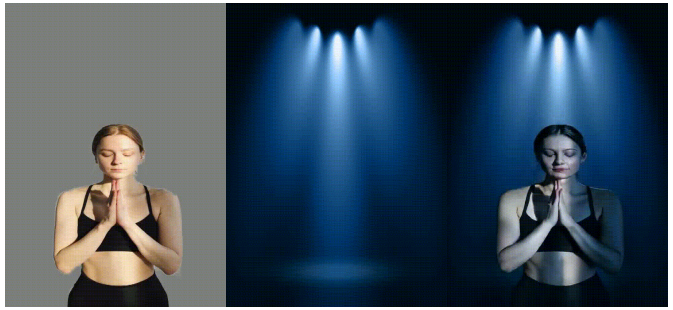


"..., warm-gleaming Fresnel lamps, ···"

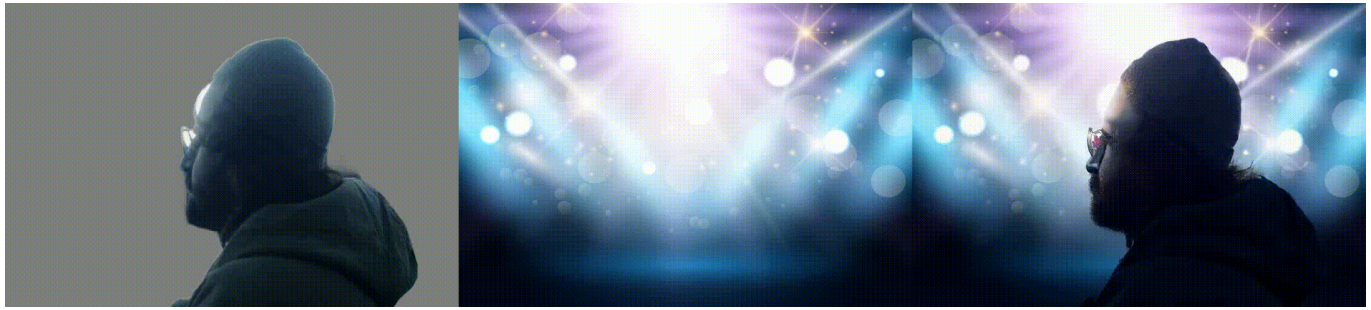


"A woman with short dark hair, ···, focused pure white beam, inky black void"

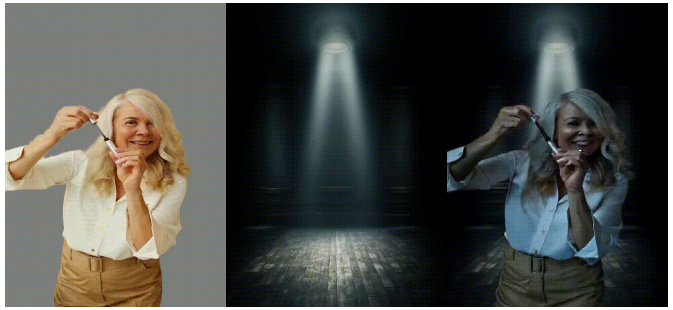
"..., cool-blue spotlights, ···"



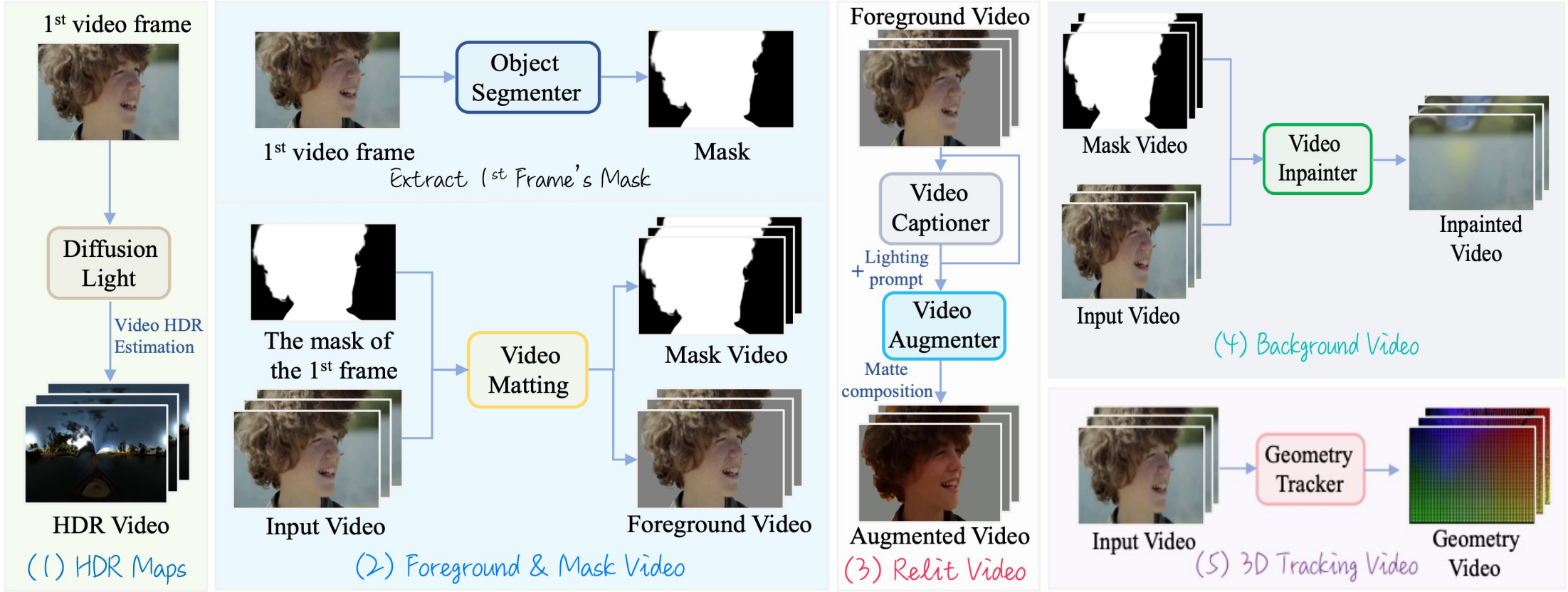
"a contemplative man in a dark beanie, ···, Soft violet and cyan spotlights"



"..., soft frosted beams, dark room, ···"



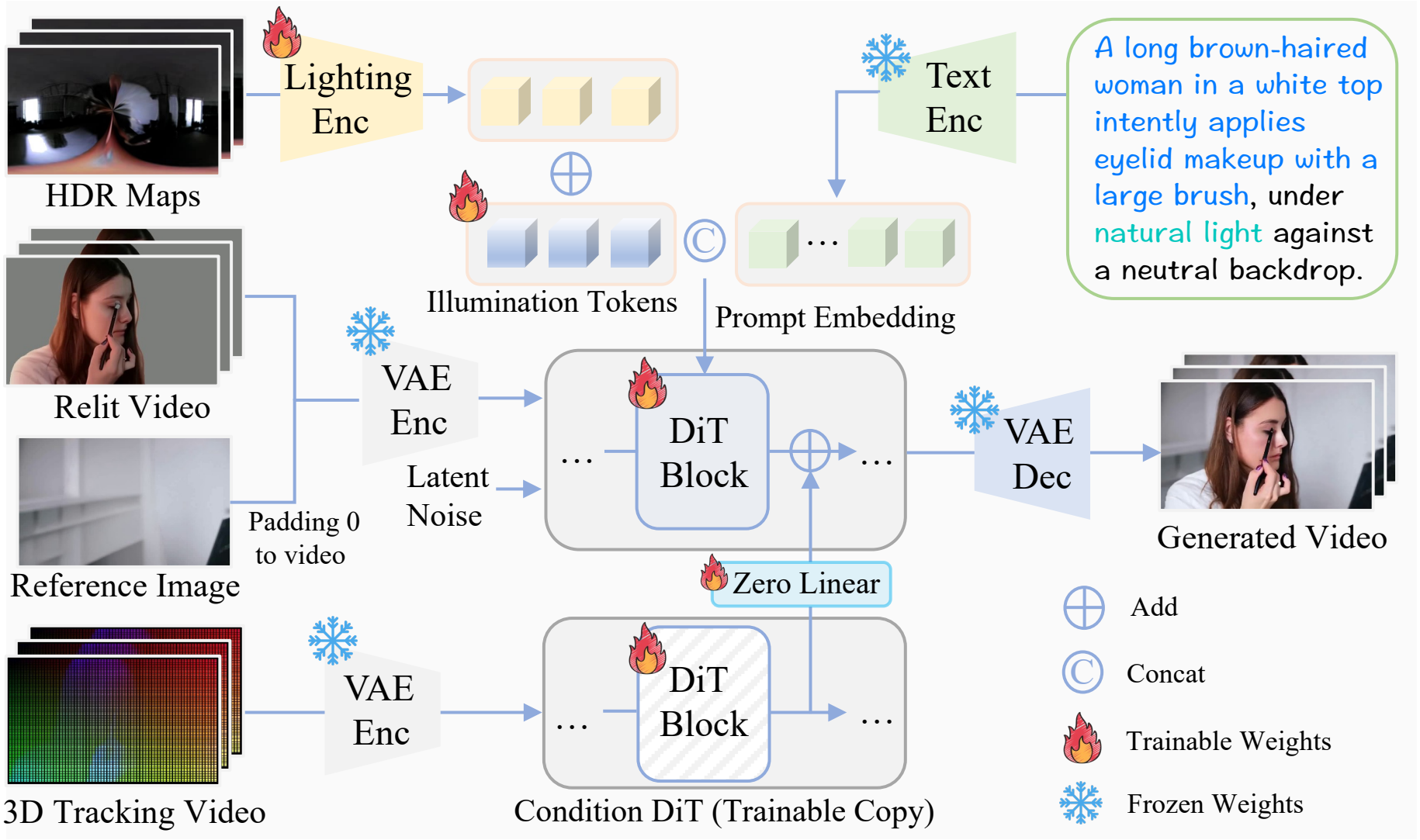
Method (IllumiPipe)



Object Segmenter: Grounded SAM-2 [42]
Video Captioner: CogVLM2-Video-LLaMA3-Chat [46]
Video Inpainter: DiffEraser [44]

Video Matting Model: MatAnyone [43]
Video Augmenter: Light-A-Video [1]
Geometry Tracker: SpatialTracker [45]

Method (IllumiCraft Framework)



Training Details

Training objective:

$$\min_{\theta} \mathbb{E}_{z\sim E_{\mathrm{VAE}}(x),\, t,\, \epsilon\sim\mathcal{N}(0,1)} \left\| \epsilon - \epsilon_\theta\left(z_t,\, t,\, \mathcal{E}\right) \right\|_2^2, \quad \mathcal{E} = \{z_g,\, z_c,\, \mathcal{P}'\}.$$

z_t: noisy latent at diffusion step t

z_g: geometry latent

z_c: control latent

$\mathcal{P}'$: final prompt embedding

Training setup:

Data source: Pexels

DiT backbone: Wan2.1 1.3B

Dropping rate: HDR feature (50%), 3D tracking video (30%), reference image (10%)

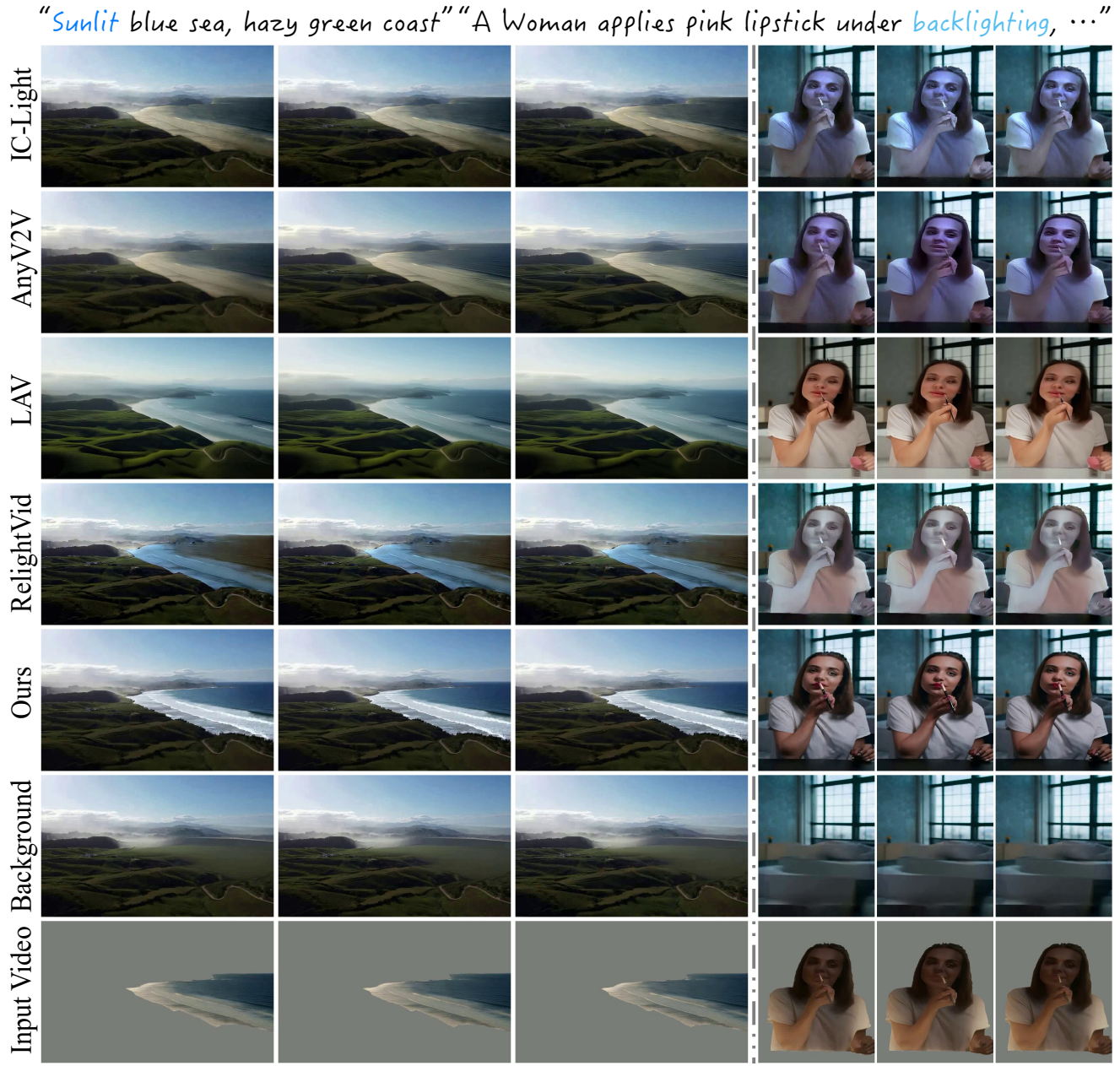
Training strategy: 4 A6000 GPUs, 3000 iterations, Adam optimizer, 4e-5 learning rate, batch size of 2

Training time: ~ 90 hours

Experiments

Visual results under the text-conditioned setting

Experiments



Visual results under the background-conditioned setting

Experiments

Benchmarking text-conditioned video relighting:

Method	FVD (↓)	LPIPS (↓)	PSNR (↑)	Text Alignment (↑)	Temporal Consistency (↑)
IC-Light [9]	4914.83	0.7330	8.55	0.3091	0.9508
AnyV2V [32] + IC-Light	3857.09	0.6979	11.12	0.2781	0.9808
Light-A-Video [1]	3946.71	0.6754	11.71	0.3020	0.9910
IllumiCraft	2186.40	0.5623	12.03	0.3342	0.9948

Benchmarking background-conditioned video relighting:

Method	FVD (↓)	LPIPS (↓)	PSNR (↑)	Text Alignment (↑)	Temporal Consistency (↑)
IC-Light [9]	2175.97	0.3049	17.20	0.3037	0.9795
AnyV2V [32] + IC-Light	1901.41	0.3447	17.98	0.3021	0.9854
Light-A-Video [1]	1704.63	0.3834	15.64	0.3266	0.9912
RelightVid* [8]	1492.18	0.2989	17.19	0.3055	0.9858
IllumiCraft*	1011.08	0.2232	19.78	0.3283	0.9932
IllumiCraft	1072.38	0.2592	19.44	0.3292	0.9945

∗ denotes results evaluated with the first 16 frames,
49 frames are adopted by default.

Experiments (Ablation Studies)

Impact of using illumination and geometry guidance during training:

Guidance	FVD (↓)	LPIPS (↓)	PSNR (↑)	Text Alignment (↑)	Temporal Consistency (↑)
I	1305.45	0.2816	18.28	0.3211	0.9864
I + G	1072.38	0.2592	19.44	0.3292	0.9945

I: illumination guidance G: geometry guidance

Effect of dropping X_{hdr} (text-only):

Possibility	FVD (↓)	TA (↑)	TC (↑)
40%	2172.35	0.3325	0.9942
50%	2186.40	0.3342	0.9948
60%	2123.23	0.3312	0.9932
70%	2138.35	0.3301	0.9923

Effect of dropping X_{hdr} (background):

Possibility	FVD (↓)	TA (↑)	TC (↑)
40%	1048.24	0.3265	0.9926
50%	1072.38	0.3292	0.9945
60%	1065.21	0.3277	0.9937
70%	1051.14	0.3228	0.9910

X_{hdr} : the feature of HDR maps

Experiments (Ablation Studies)

Impact of dropping 3D tracking videos (text-only)

Possibility	FVD (↓)	TA (↑)	TC (↑)
10%	2285.32	0.3303	0.9893
20%	2251.21	0.3332	0.9939
30%	2186.40	0.3342	0.9948
40%	2234.35	0.3325	0.9915

Impact of dropping 3D tracking videos (background)

Possibility	FVD (↓)	TA (↑)	TC (↑)
10%	1154.32	0.3231	0.9902
20%	1102.21	0.3273	0.9935
30%	1072.38	0.3292	0.9945
40%	1098.35	0.3278	0.9937

Impact of dropping the reference image (text-only)

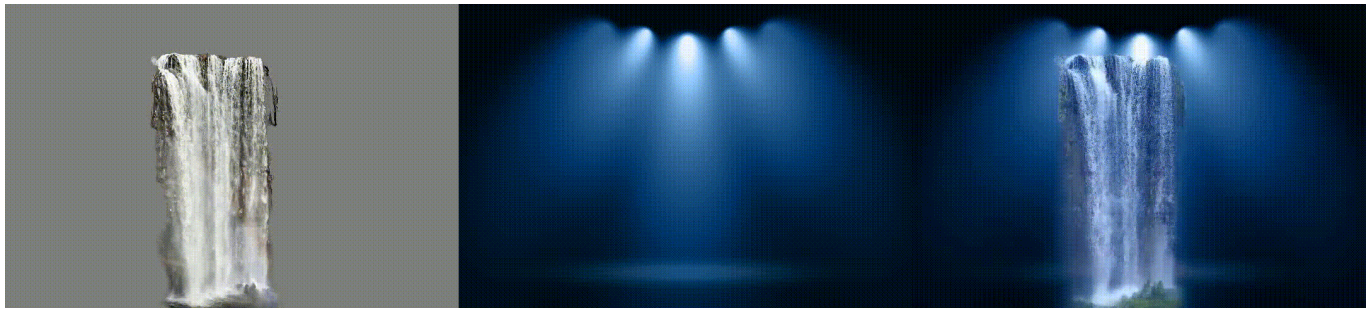
Possibility	FVD (↓)	TA (↑)	TC (↑)
5%	2232.46	0.3331	0.9939
10%	2186.40	0.3342	0.9948
20%	2175.23	0.3341	0.9943
30%	2158.32	0.3338	0.9941

Impact of dropping the reference image (background)

Possibility	FVD (↓)	TA (↑)	TC (↑)
5%	1065.83	0.3284	0.9941
10%	1072.38	0.3292	0.9945
20%	1105.28	0.3275	0.9928
30%	1127.32	0.3269	0.9925

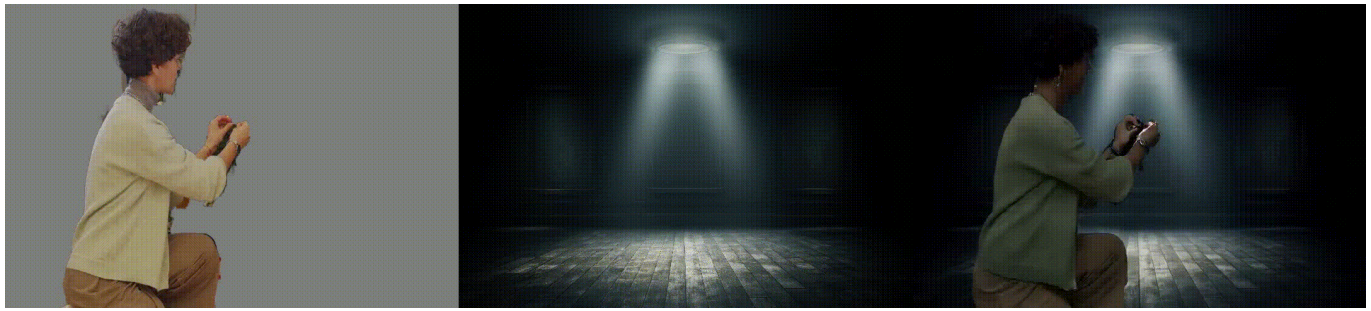
Additional Results

"A majestic waterfall, ⋯, cool-blue spotlights, ⋯"

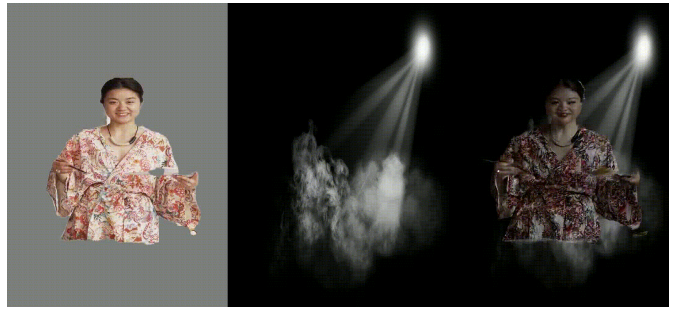


"..., soft-edged white beams, ⋯"

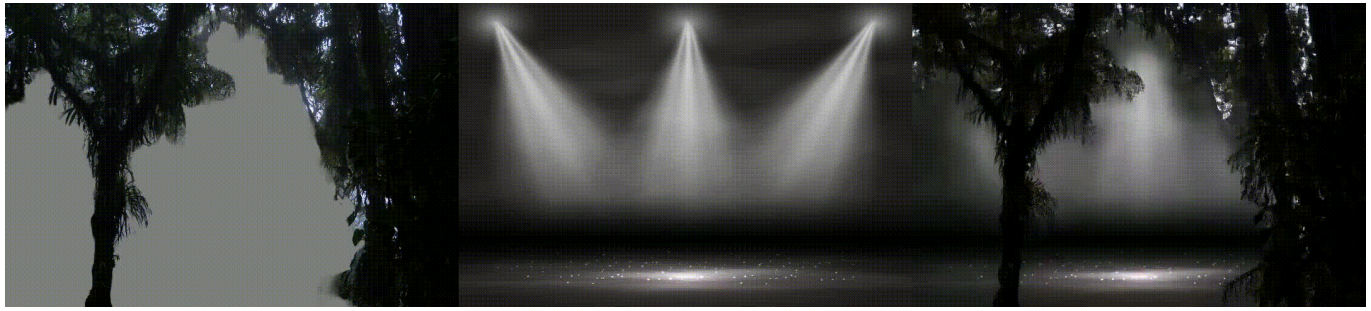
"A young woman in a light green sweater, ⋯, soft frosted beams, a dark room"



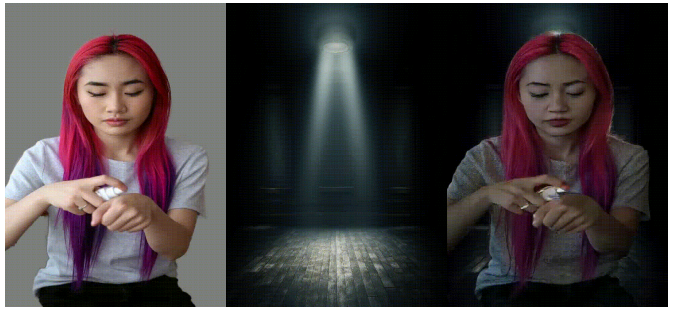
"..., luminous beam, thick smoke, ⋯"



"A serene, mystical forest shrouded in fog, ⋯, narrow white spotlight, intimate vibe"

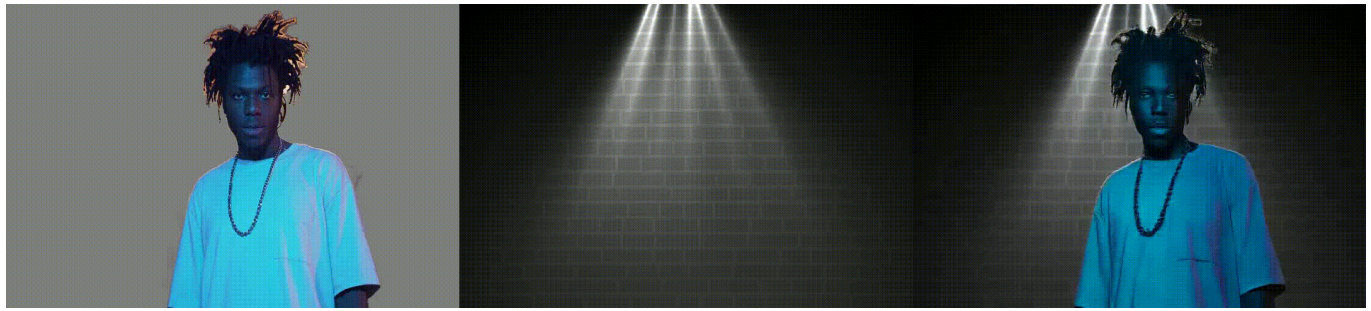


"..., soft frosted beams, dark room, ⋯"



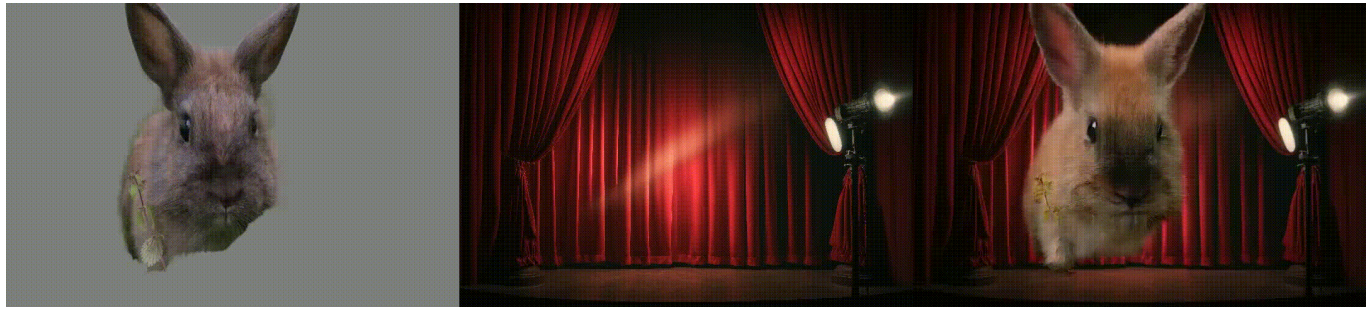
Additional Results

"A young man with a serious expression, ···, sharp shafts of light, urban spotlight"

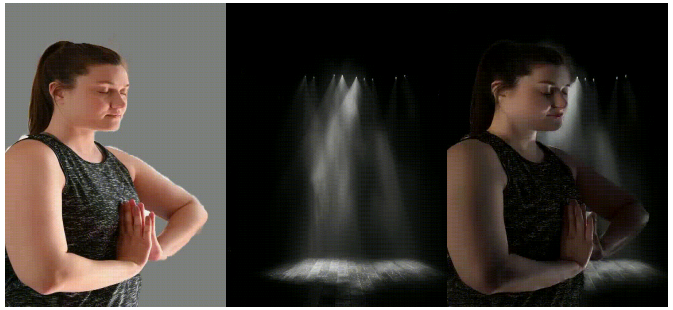


"..., luminous beam, thick smoke, ···"

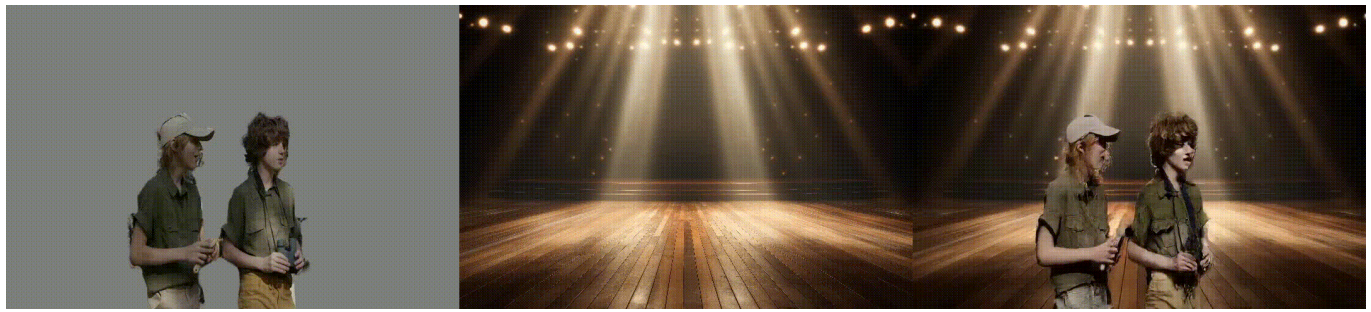
"A small, attentive rabbit, warm-gleaming Fresnel lamps, red velvet curtains"



"..., milky spotlights, grey haze, ···"



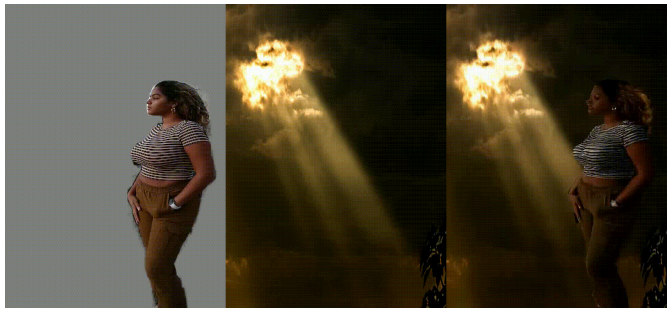
"Two young individuals, dressed in outdoor attire, ···, warm amber spotlights"



"..., white beams of daylight, ···"

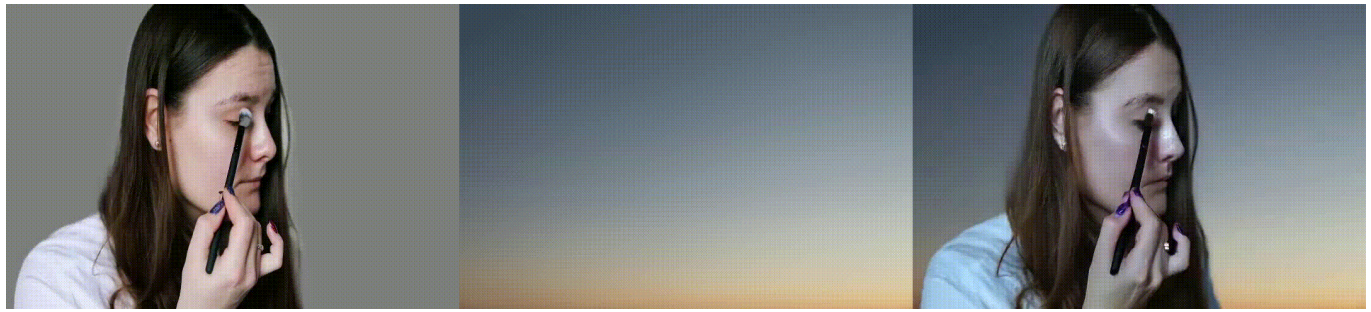
Additional Results

"A single purple iris flower, …, focused pure white beam, inky black void, …" "…, Warm golden crepuscular rays, …"



"A hooded man in hard-edged luminous spotlight over a dark floor, suspenseful stage mood" "…, pastel-blue and lilac LED, …"

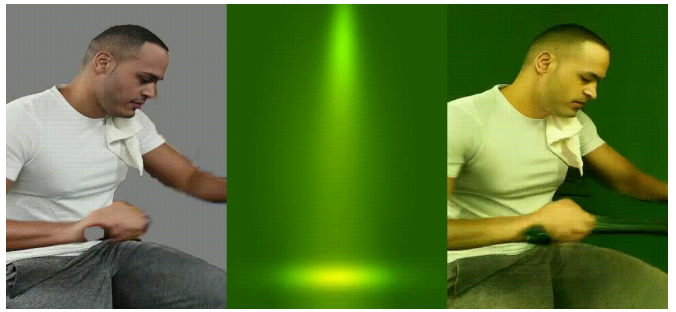
"A young woman with long brown hair, …, twilight gradient washes, …" "…, crisp ice-white beams, …"



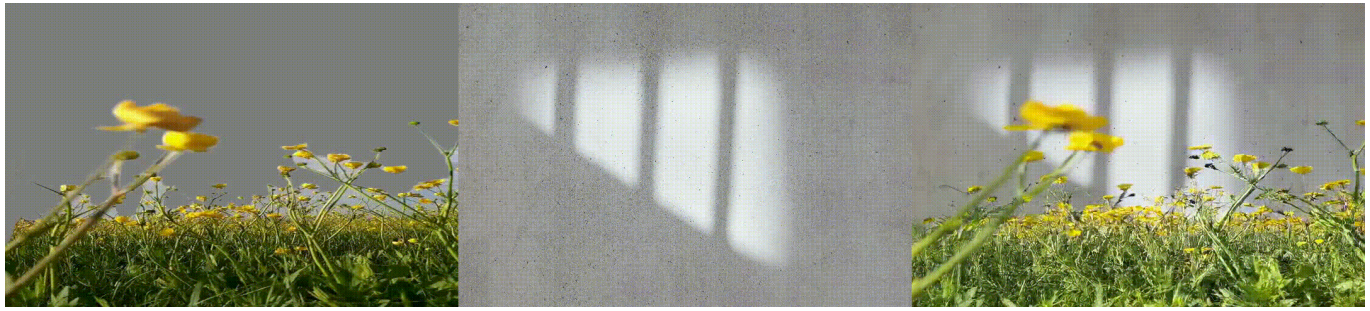
Additional Results

"A young black woman, …, cool white beam, dramatic noir-style atmosphere"

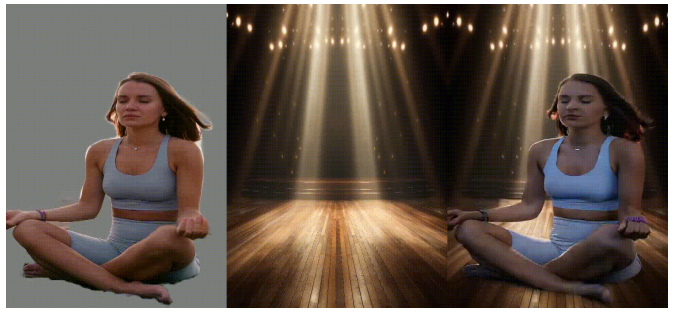
"…, vivid chartreuse spotlight, …"



"yellow flowers in full bloom, diffused daylight, a calm, minimalist feel"



"…, warm amber spotlights, …"



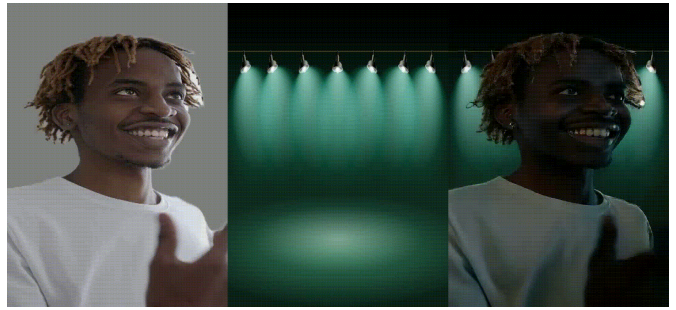
"a dense forest of coniferous trees dusted with snow, crisp ice-white beams, … "

"…, white fluorescent diffused light, …"

Additional Results

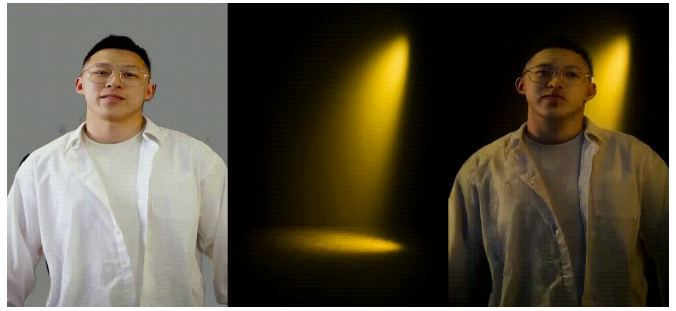
"A stationary airplane wing, …, crisp bright spotlights, …"

"…, teal-tinted light, …"

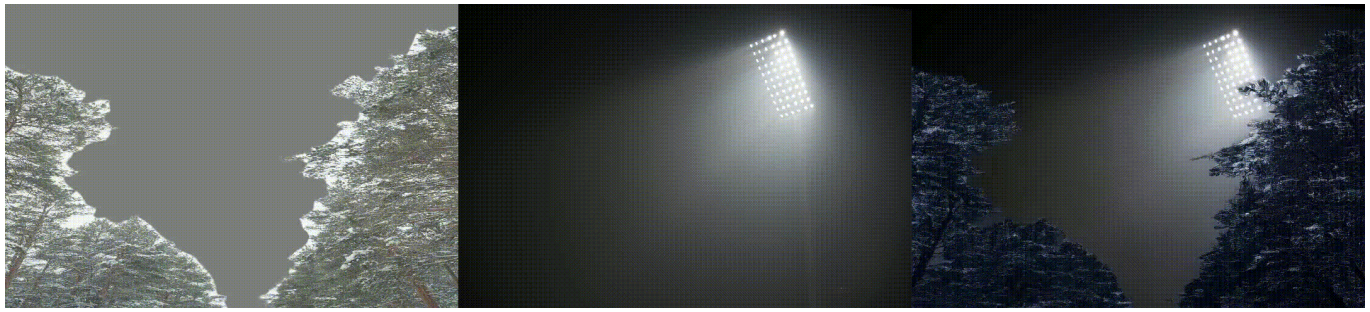


"A dark car buried in deep snow, …, warm-white wall sconces, …"

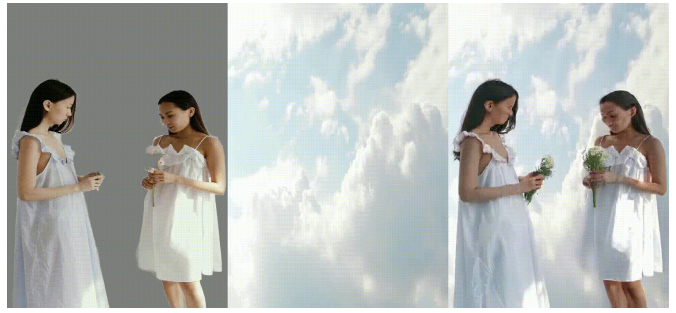
"…, warm gold spotlight, …"



"Snow-covered pine trees, …, towering LED floodlight, stark white illumination"

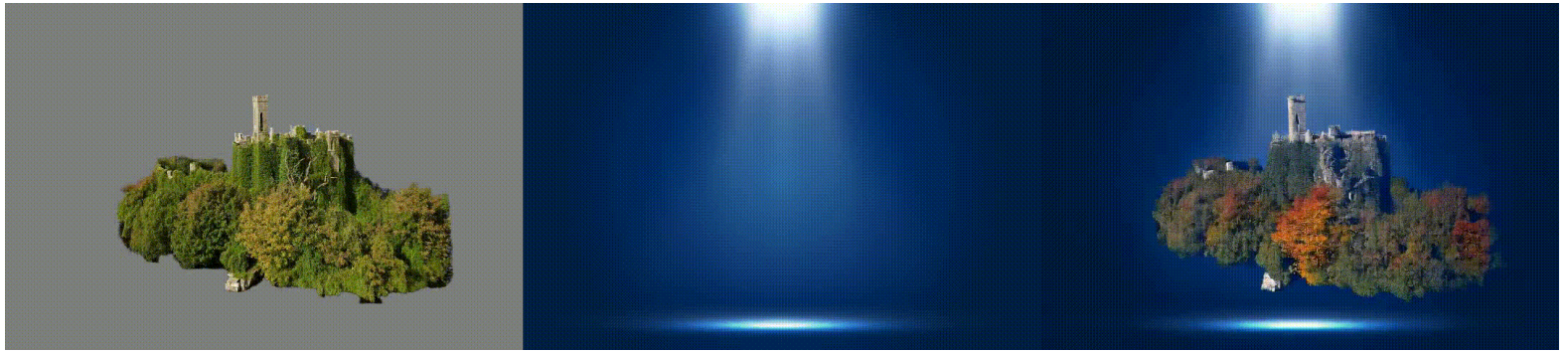


"…, diffused daylight, …"



Failure Cases

"An ancient stone castle on an island, ···, bright chalky spotlight in misty blue haze"



"branch of an apple tree in full bloom, ···, towering LED floodlight, stark snowy illumination"

Thank You !!!