Diffusion Guided Chain-of-Vision for Large Autoregressive Vision Models

Supplementary Material

Appendix

A. DDIM Inversion Implementation

We present a three-stage framework for DDIM inversion [7, 12] and latent space manipulation, as detailed in Algorithm 1-3.

A.1. DDIM Inversion

Algorithm 1 performs the deterministic inversion of an input image into the noise latent space. Given an input image I, we first encode it into a clean latent representation z_0 using a VAE [5] encoder E. The inversion process then iteratively transforms this clean latent through T time steps using the DDIM [9] formulation. At each step t, the noise predictor ϵ_θ estimates the noise component ϵ_t from the current latent z_t and time step t. The coefficient η_t is computed based on the noise schedule $\{\bar{\alpha}_t\}$, and the latent is updated according to the DDIM inversion equation.

A.2. Spherical Linear Interpolation

Algorithm 2 performs interpolation in the noise space using spherical linear interpolation (slerp) [8]. Given source and target noise latents z_T^0 and z_T^1, we first compute the angular distance ϕ between them. We then generate L intermediate latents by uniformly sampling along the geodesic (great circle) connecting the endpoints. For each interpolation factor α_k, the corresponding latent $z_T^{\alpha_k}$ is computed with the SLERP formula, which ensures constant angular velocity and preserves the latent-space geometry. This yields a smooth sequence Z bridging the two noise latents.

A.3. Deterministic Denoising

Algorithm 3 performs deterministic denoising of the interpolated latent sequence back to the image space. For each interpolated noise latent $z_T^{\alpha_k}$ in sequence Z, we execute the reverse DDIM process from time step T down to 1. At each denoising step, the noise predictor ϵ_θ estimates the noise component, and the latent is updated using the DDIM denoising step. After completing the denoising trajectory, the clean latent z_0 is decoded into the final image $\mathcal{I}_{\alpha_k}$ using the VAE decoder D.

Algorithm 1 Stage-1: DDIM Inversion

Input: Image I, VAE encoder E, noise predictor ϵ_θ, step T, schedule $\{\bar{\alpha}_t\}_0^T$.

$z_0 \leftarrow E(I)$ // Encode to clean latent
for $t = 0$ to $T-1$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(z_t, t)$ // Predict noise
  $\eta_t \leftarrow \sqrt{\frac{1-\bar{\alpha}_{t+1}}{\bar{\alpha}_{t+1}}} - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}$ // Compute coefficient
  $z_{t+1} \leftarrow \sqrt{\bar{\alpha}_{t+1}} \left(\frac{z_t}{\sqrt{\bar{\alpha}_t}} + \eta_t \cdot \epsilon_t \right)$ // Update latent
end for
return z_T

Algorithm 2 Stage-2: Latent Interpolation in Noise Space

Input: Source noise latent z_T^0, target noise latent z_T^1, interpolation length $L \in \mathbb{N}$

$\phi \leftarrow \arccos\left(\frac{(z_T^0)^\top z_T^1}{\|z_T^0\| \, \|z_T^1\|} \right)$ // Angle between noise latents
$Z \leftarrow \emptyset$ // Initialize sequence
for $k = 1$ to L **do**
  $\alpha_k \leftarrow \frac{k}{L+1}$ // Uniform sampling in [0,1]
  $z_T^{\alpha_k} \leftarrow \frac{\sin((1-\alpha_k)\phi)}{\sin\phi} z_T^0 + \frac{\sin(\alpha_k \phi)}{\sin\phi} z_T^1$ // Interpolate
  $Z \leftarrow Z \cup \{z_T^{\alpha_k}\}$
end for
return Z

Algorithm 3 Stage-3: DDIM Denoising

Input: Interpolated noise sequence $Z = \{z_T^{\alpha_k}\}_{k=1}^L$, VAE decoder D, noise predictor ϵ_θ, step T, schedule $\{\bar{\alpha}_t\}_{t=0}^T$

$I \leftarrow \emptyset$
for $k = 1$ to L **do**
  $z_T \leftarrow z_T^{\alpha_k}$
  for $t = T$ down to 1 **do**
    $\epsilon_t \leftarrow \epsilon_\theta(z_t, t)$ // Predict noise
    $\eta_t \leftarrow \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}} - \sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}}$ // Compute coefficient
    $z_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z_t}{\sqrt{\bar{\alpha}_t}} - \eta_t \cdot \epsilon_t \right)$ // Update latent
  end for
  $\mathcal{I}_{\alpha_k} \leftarrow D(z_0)$ // Decode to image
  $I \leftarrow I \cup \{\mathcal{I}_{\alpha_k}\}$
end for
return I

B. Comparison Results

B.1. Spherical Linear Interpolation

We evaluate the effect of spherical linear interpolation in the latent space; results are summarized in Tables S1 and S2.

Table S1. Effect of spherical linear interpolation on image segmentation with LVM-7B.

Method	MS-COCO	
	IoU↑	P-ACC↑
CoV *w/.* Linear	0.498	0.606
CoV *w/.* Spherical Linear	**0.510**	**0.619**

Table S2. Effect of spherical linear interpolation on depth estimation with LVM-7B.

Method	ImageNet-1k	
	A.Rel↓	S.Rel↓
CoV *w/.* Linear	0.324	0.072
CoV *w/.* Spherical Linear	**0.312**	**0.060**

B.2. Inference Time

As shown in Table S3, we report inference time as the average latency to generate one sentence output with LVM-7B. All measurements are taken on an NVIDIA H20 GPU.

Table S3. Inference time comparison.

Method	Inference time (ms) ↓
LVM-7B *w/o.* interp.	504
CoF [13]	824
CoV *w/.* 1-interp.	842
CoV *w/.* 2-interp.	1130

C. Implementation Details

C.1. Model Architecture

We employ three LVMs of different sizes, all based on LLaMA [10]. Architectural configurations are listed in Table S4.

Table S4. Architectural hyperparameters for CoV.

Model	Hidden dim.	MLP dim.	#heads	#layers
LVM-300M [4]	1024	2688	8	22
LVM-1B [4]	2048	5504	16	22
LVM-7B [2]	4096	11008	32	32

C.2. Training details

Following prior LVMs [2, 4], we train our models with the AdamW optimizer. The full training configurations are provided in Tables S5 and S6. All models are trained on NVIDIA H20 GPUs.

Table S5. Training hyperparameters for CoV on LVM-300M and LVM-1B.

Config	Value
Optimizer	AdamW
Learning rate	1e-3
Epochs	50
Batch size	224 visual sentences (172,032 tokens)
Weight decay	0.01
Betas	(0.9, 0.95)
LR scheduler	One-cycle cosine
Warm-up steps	5% of total steps
LoRA alpha	64
LoRA rank (r)	32
LoRA dropout	0.1
Context length	2048

Table S6. Training hyperparameters for CoV on LVM-7B.

Config	Value
Optimizer	AdamW
Learning rate	1e-3
Epochs	20
Batch size	224 visual sentences (172,032 tokens)
Weight decay	0.01
Betas	(0.9, 0.95)
LR scheduler	One-cycle cosine
Warm-up steps	10% of total steps
LoRA alpha	64
LoRA rank (r)	32
LoRA dropout	0.1
Context length	4096

D. Additional Results

This appendix presents additional Chain-of-Vision results across multiple vision tasks.

D.1. Image Segmentation

Additional qualitative results for image segmentation on the MS COCO [6] dataset are shown in Figure S1.

D.2. Human Pose Estimation

Additional qualitative results for human pose estimation on the COCO-Pose [1] dataset are shown in Figure S2.

D.3. Image Colorization

Additional qualitative results for image colorization on the ImageNet-1K [3] dataset are shown in Figure S3.

D.4. Surface Normal Estimation

Additional qualitative results for surface normal estimation on the ImageNet-1K [3] dataset are shown in Figure S4.

D.5. Edge Detection

Additional qualitative results for edge detection on the ImageNet-1K [3] dataset are shown in Figure S5.

D.6. Depth Estimation

Additional qualitative results for depth estimation on the ImageNet-1K [3] dataset are shown in Figure S6.

D.7. Low-Light Enhancement

Additional qualitative results for low-light enhancement on the LoL [11] dataset are shown in Figure S7.

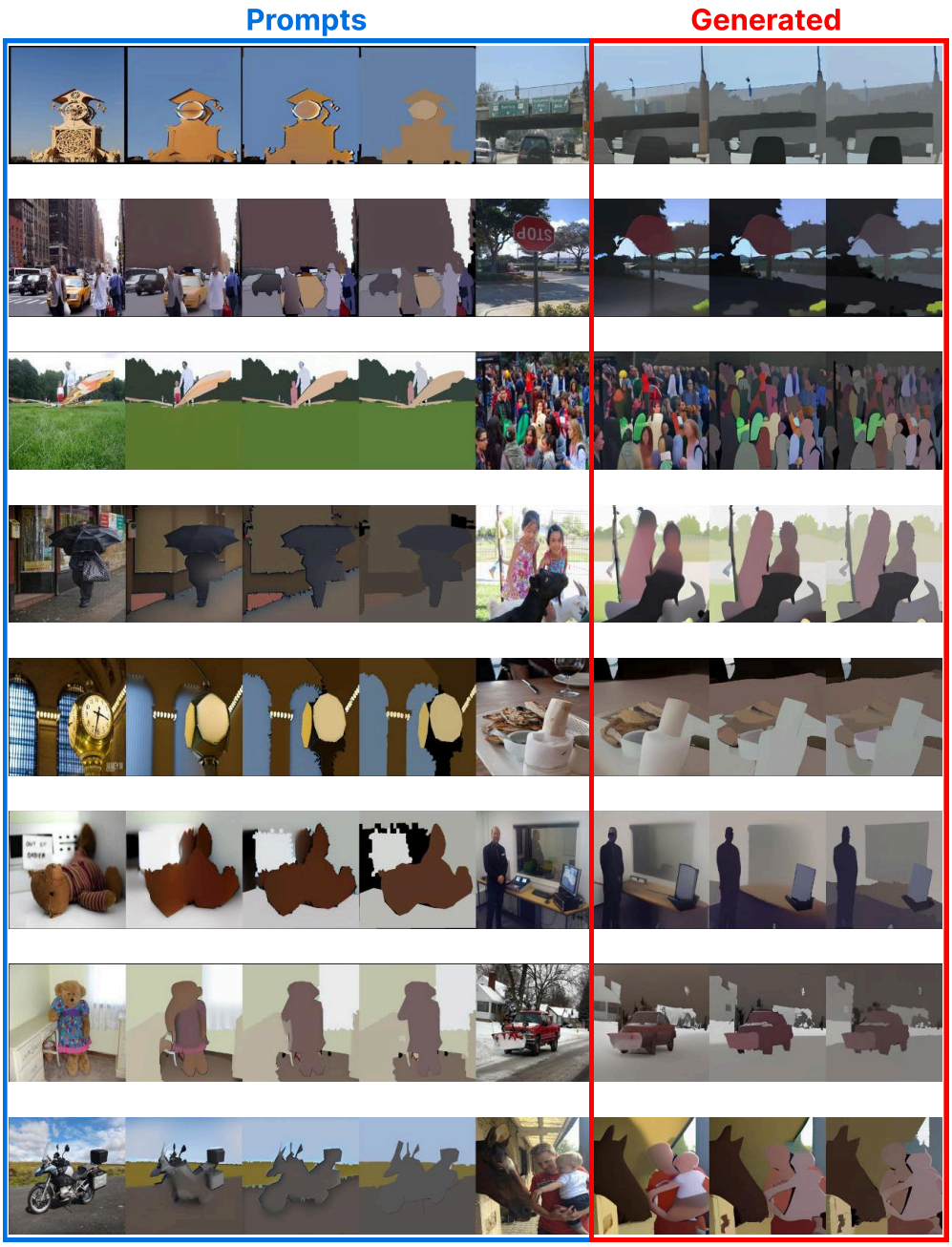

Figure S1. **Image segmentation.** Additional qualitative results.

Figure S2. **Human pose estimation.** Additional qualitative results.

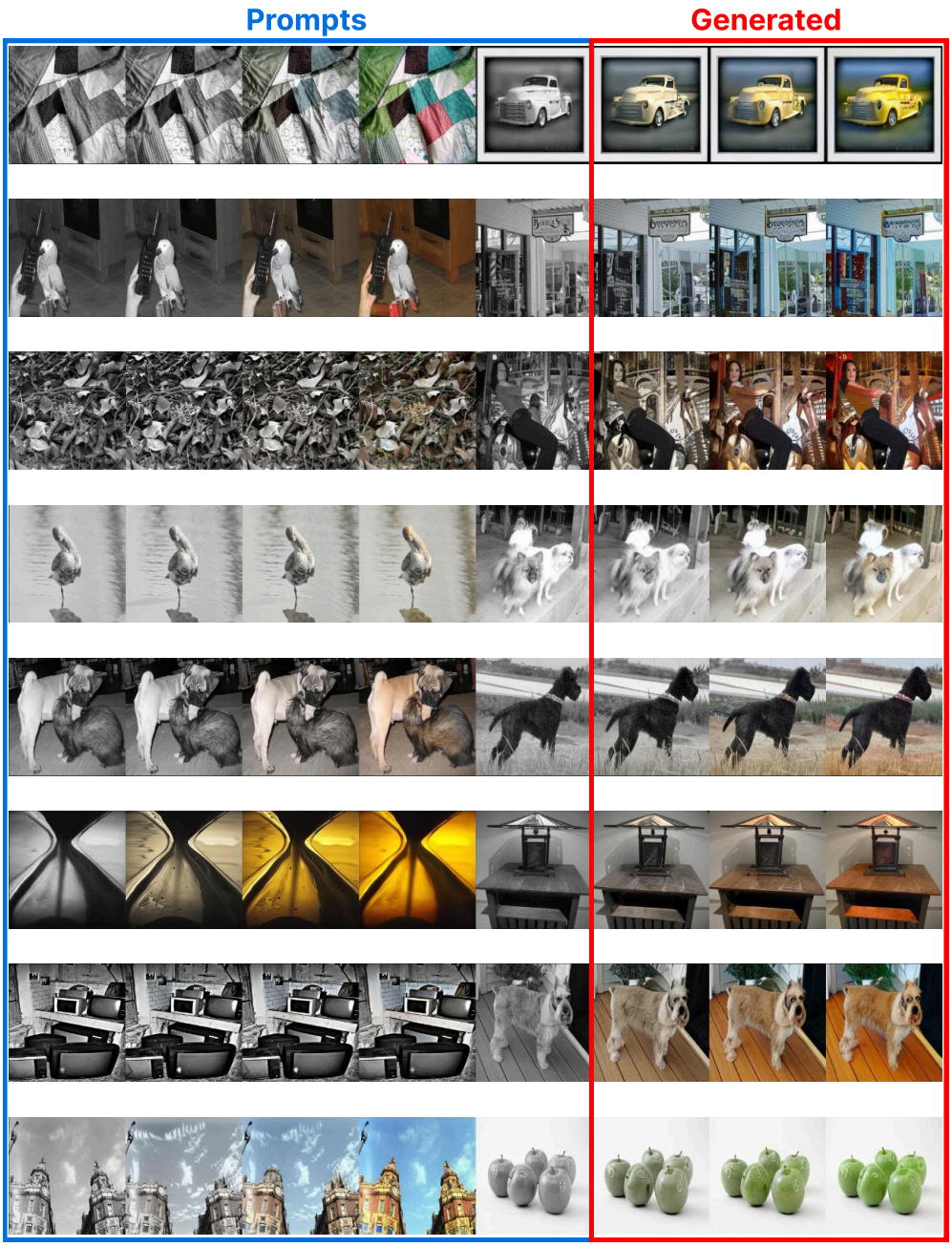

Figure S3. **Image colorization.** Additional qualitative results.

Figure S4. **Surface normal estimation.** Additional qualitative results on the ImageNet-1K dataset.

Figure S5. **Edge detection.** Additional qualitative results.

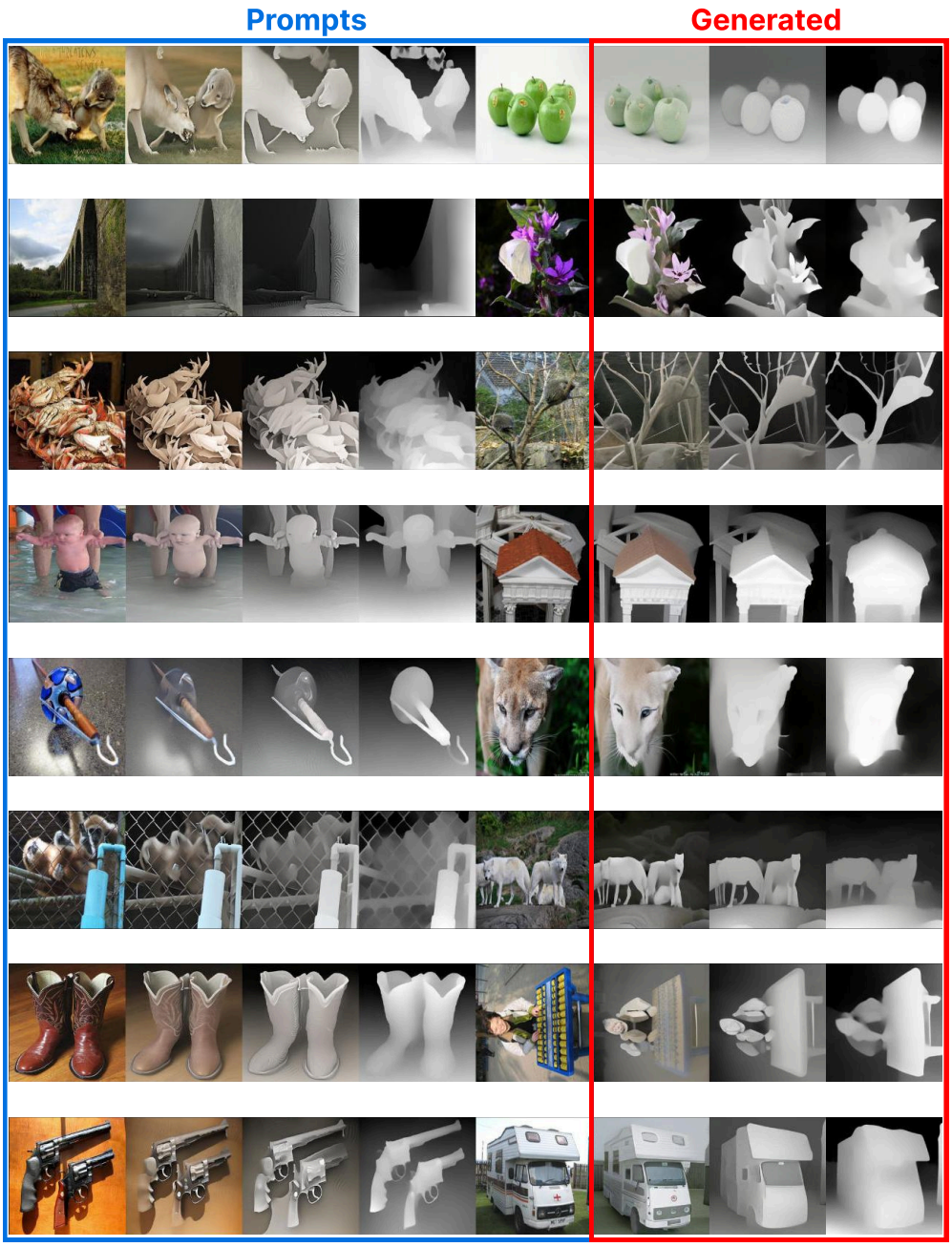

Figure S6. **Depth estimation.** Additional qualitative results on the ImageNet-1K dataset.

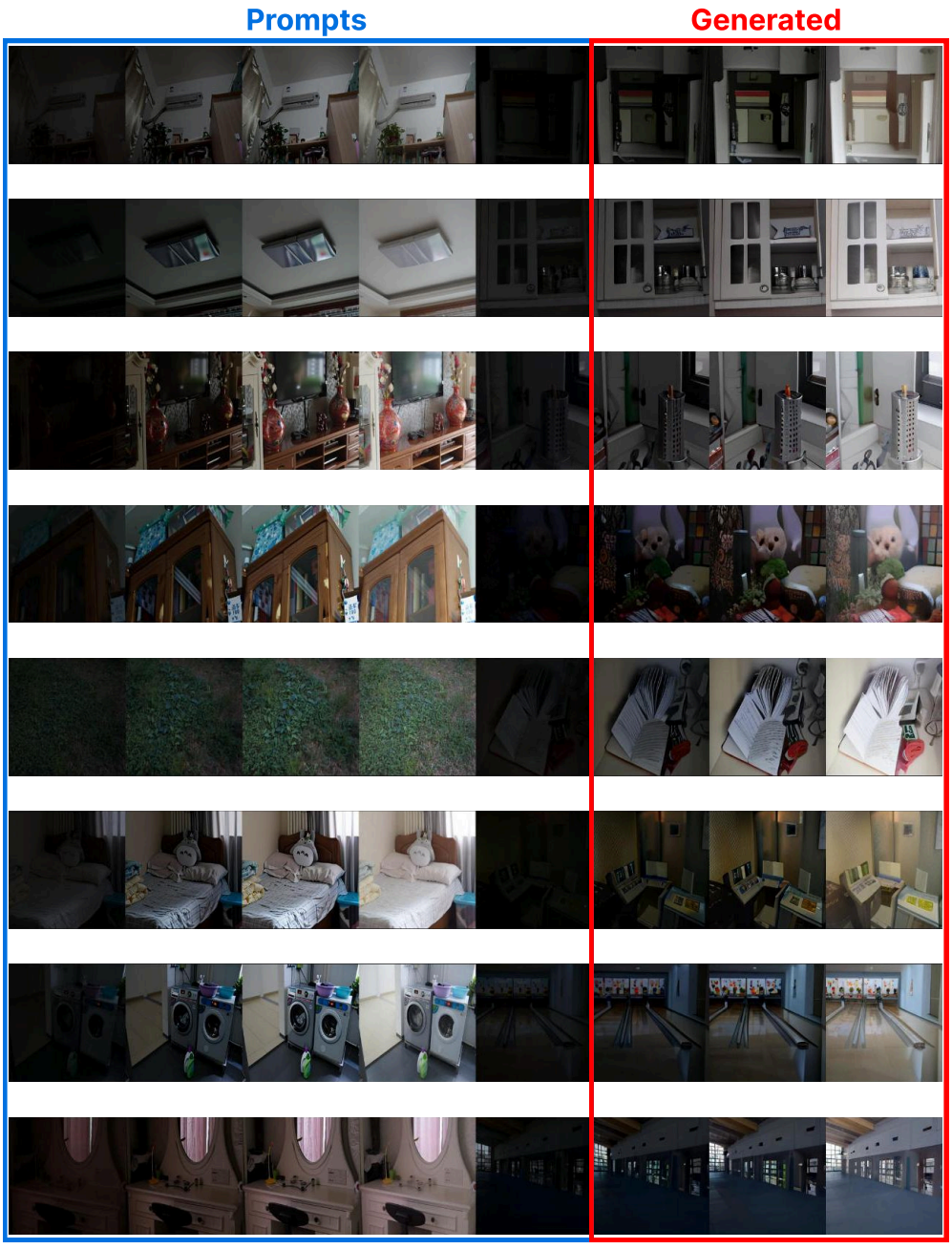

Figure S7. **Low-light enhancement.** Additional qualitative results.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[2] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Jianyuan Guo, Zhiwei Hao, Chengcheng Wang, Yehui Tang, Han Wu, Han Hu, Kai Han, and Chang Xu. Data-efficient large vision models through sequential autoregression. *arXiv preprint arXiv:2402.04841*, 2024. 2

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2

[7] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 1

[8] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985. 1

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[11] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 3

[12] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2024. 1

[13] Jiyang Zheng, Jialiang Shen, Yu Yao, Min Wang, Yang Yang, Dadong Wang, and Tongliang Liu. Chain-of-focus prompting: Leveraging sequential visual cues to prompt large autoregressive vision models. In *The Thirteenth International Conference on Learning Representations*. 2