Cross-Resolution Diffusion Models via Network Pruning

Supplementary Material

6. Block-wise Pruning Ratio Configurations

As discussed in Section 3.1, the UNet architecture comprises downsampling, middle, and upsampling blocks, which differ in redundancy and tolerance to parameter removal. This is further supported by our pruning ratio search experiments across multiple diffusion model families and sampling resolutions, with the resulting block-wise configurations summarized in Table 6.

The empirical results consistently show that the optimal pruning ratios vary across the three block groups in SDXL, SD1.5, and SD2.1, and this difference remains stable when the generation resolution changes. These observations indicate that each block group contributes to synthesis in a structurally differentiated manner and therefore exhibits distinct pruning sensitivity. Applying a uniform pruning ratio across all blocks either disrupts global structural composition or suppresses fine-grained details. In contrast, assigning pruning ratios separately to the downsampling, middle, and upsampling blocks maintains texture fidelity.

Taken together, these findings directly support our *Block-wise Pruning Ratio* Strategy in Section 3.3.

Table 6. Block-wise pruning ratio configurations across different models and resolutions, showing the distinct ratio allocations for the downsampling, middle, and upsampling blocks (* indicates the model's default resolution).

Model	Resolution	Ratios (Down/Middle/Up)
SDXL	$1024 \times 1024^*$	0.295 / 0.194 / 0.236
	512×512	0.397 / 0.434 / 0.387
	400×560	0.482 / 0.396 / 0.469
	480×360	0.434 / 0.428 / 0.355
	1536×1536	0.300 / 0.343 / 0.300
SD1.5	$512 \times 512^*$	0.433 / 0.345 / 0.300
	400×560	0.319 / 0.240 / 0.192
	480×360	0.467 / 0.363 / 0.196
	768×768	0.185 / 0.445 / 0.100
SD2.1	$512 \times 512^*$	0.623 / 0.259 / 0.115
	400×560	0.534 / 0.534 / 0.169
	480×360	0.651 / 0.138 / 0.271
	768×768	0.277 / 0.206 / 0.313

7. Full Ablation Study of POA

To more comprehensively illustrate the effect of the *pruned output amplification* (POA) mechanism, we provide the full ablation results across models and resolutions in Table 7, which were omitted from the main paper due to space constraints.

This output-level refinement consistently improves generative quality across architectures and resolutions. As shown in Table 7, the refined models achieve stronger semantic consistency and perceptual fidelity, reflected in higher ImageReward and PickScore values compared with pure-pruned baselines. These results indicate that POA functions as a corrective steering mechanism that stabilizes the denoising process and reinforces the desirable generative tendencies of the pruned model while reducing residual artifact-related signals inherited from the dense model.

Discussions on Aesthetic Score. We observe that Aesthetic Scores may occasionally decrease after applying POA, even as ImageReward, CLIP, and PickScore show consistent improvements. This phenomenon occurs because the Aesthetic Score is particularly sensitive to variations in local texture and stylistic details. By pushing the output further along the pruned direction (k>1), POA naturally moderates certain fine-grained textural components that tend to exhibit instability at unseen resolutions. This moderation results in smoother and more structurally coherent outputs, which may not align perfectly with the Aesthetic Score's emphasis on textural richness. However, the consistent gains observed in ImageReward and PickScore metrics demonstrate improved semantic alignment, enhanced realism, and superior overall visual coherence, thereby validating the effectiveness of POA.

8. Simulated Annealing (SA) Algorithm

Algorithm 1 summarizes the simulated annealing (SA) routine used to search for the optimal pruning ratio configuration $\mathbf{r} = r_{\text{down}}, r_{\text{mid}}, r_{\text{up}}$. The hyperparameters include the initial temperature T_{init}, cooling rate α, iteration budget N_{iter}, a set of candidate seeds S_{seeds}, and a restart limit R_{max}. Starting from the best candidate in the initial seed set, the algorithm iteratively samples neighboring configurations and accepts them based on the standard SA criterion, allowing occasional uphill moves to escape local minima. A lightweight reheating and restart mechanism is incorporated to prevent stagnation and maintain exploration when the search plateaus. This SA variant provides a simple and robust way to obtain near-optimal ratio configurations without exhaustive search, and the resulting best state S_{best} serves directly as the optimal pruning–ratio configuration $\mathbf{r}$.

Table 7. Performance comparison of *pruned* models (the model pruned block-wisely without pruned output amplification) and *CR-Diff* (with pruned output amplification) across default and unseen resolutions. Resolutions are reported as height × width, where resolutions marked with * denote the native (default) setting of the model. **Bold** values indicate that CR-Diff outperforms the pruned baseline. These results verify that POA provides a stable refinement effect that generalizes across models and resolutions.

Model	Resolution	FID ↓		CLIP ↑		ImageReward ↑		PickScore ↑		Aesthetic Score ↑	
		pruned	CR-Diff	pruned	CR-Diff	pruned	CR-Diff	pruned	CR-Diff	pruned	CR-Diff
SDXL	1024 × 1024*	33.397	33.562	0.322	0.322	0.834	**0.946**	22.594	**22.639**	6.058	**6.106**
	512 × 512	40.068	**37.918**	0.320	**0.321**	0.530	**0.735**	22.100	**22.140**	5.508	**5.525**
	400 × 560	43.348	**36.688**	0.308	**0.311**	-0.272	**0.092**	20.948	**21.074**	4.752	4.672
	480 × 360	56.182	**46.040**	0.301	**0.307**	-0.516	**-0.099**	20.636	**20.956**	4.472	**4.644**
	1536 × 1536	39.362	40.380	0.312	0.312	0.108	**0.208**	21.394	**21.399**	5.806	**5.855**
SD1.5	512 × 512*	39.563	**37.773**	0.313	**0.314**	0.059	**0.203**	21.376	**21.377**	5.265	5.233
	400 × 560	40.188	**39.291**	0.309	**0.310**	-0.004	**0.151**	21.143	**21.188**	4.785	4.779
	480 × 360	39.774	**37.634**	0.305	**0.307**	-0.190	**-0.026**	20.931	**20.944**	4.848	4.819
	768 × 768	39.084	**38.452**	0.314	**0.315**	-0.063	**0.059**	21.190	**21.232**	5.389	5.385
SD2.1	512 × 512*	38.799	**36.792**	0.306	**0.309**	-0.288	**-0.052**	20.940	**20.960**	5.174	5.082
	400 × 560	38.344	**35.837**	0.301	**0.304**	-0.334	**-0.068**	20.565	20.540	4.559	4.428
	480 × 360	43.294	**41.042**	0.290	**0.294**	-0.822	**-0.561**	20.090	**20.177**	4.564	4.532
	768 × 768	35.595	**35.237**	0.317	**0.318**	0.304	**0.419**	21.497	21.451	5.429	5.339

9. Analyses on Unseen Resolutions

Beyond the detailed analysis in Section 4.2, which demonstrates consistent improvements under CR-Diff at unseen resolutions, we provide additional analyses at higher resolutions for SDXL. SDXL, natively trained at 1024×1024 with a resampler and high-resolution cross-attention, effectively internalizes dense object structures and sharp boundaries. As a result, scaling to 1536×1536 does not lead to noticeable degradation, with FID remaining low and perceptual metrics such as CLIP, PickScore, and Aesthetic Score staying stable as shown in Table 8. Notably, under this higher resolution, pruning-based CR-Diff successfully preserves SDXL's original generative characteristics.

10. Expanded Qualitative Analyses

Representative Teaser Results. In Figures 9 and 10, we present additional representative teaser examples following the style of Figure 1, further illustrating the effectiveness of CR-Diff in enhancing cross-resolution visual consistency over the dense SDXL [33].

Results on the 5K Dataset. In Figures 11, 12, and 13, we present additional results on a subset of 5K prompts sampled from the MS-COCO 2014 validation set [27], evaluated with SDXL, SD 2.1, and SD 1.5 across multiple resolutions. These examples show clear improvements in ImageReward and exhibit noticeably better structure preservation, semantic consistency, and fine-grained visual fidelity.

Extended Results for Prompt-Specific Optimization. In Figure 14, we present extended qualitative results from our prompt-specific optimization mentioned in Section 4.3, highlighting clear improvements in ImageReward and stronger prompt–detail correspondence across diverse input prompts.

Algorithm 1: Simulated Annealing for the Optimal Pruning Ratio Configuration $\mathbf{r}$

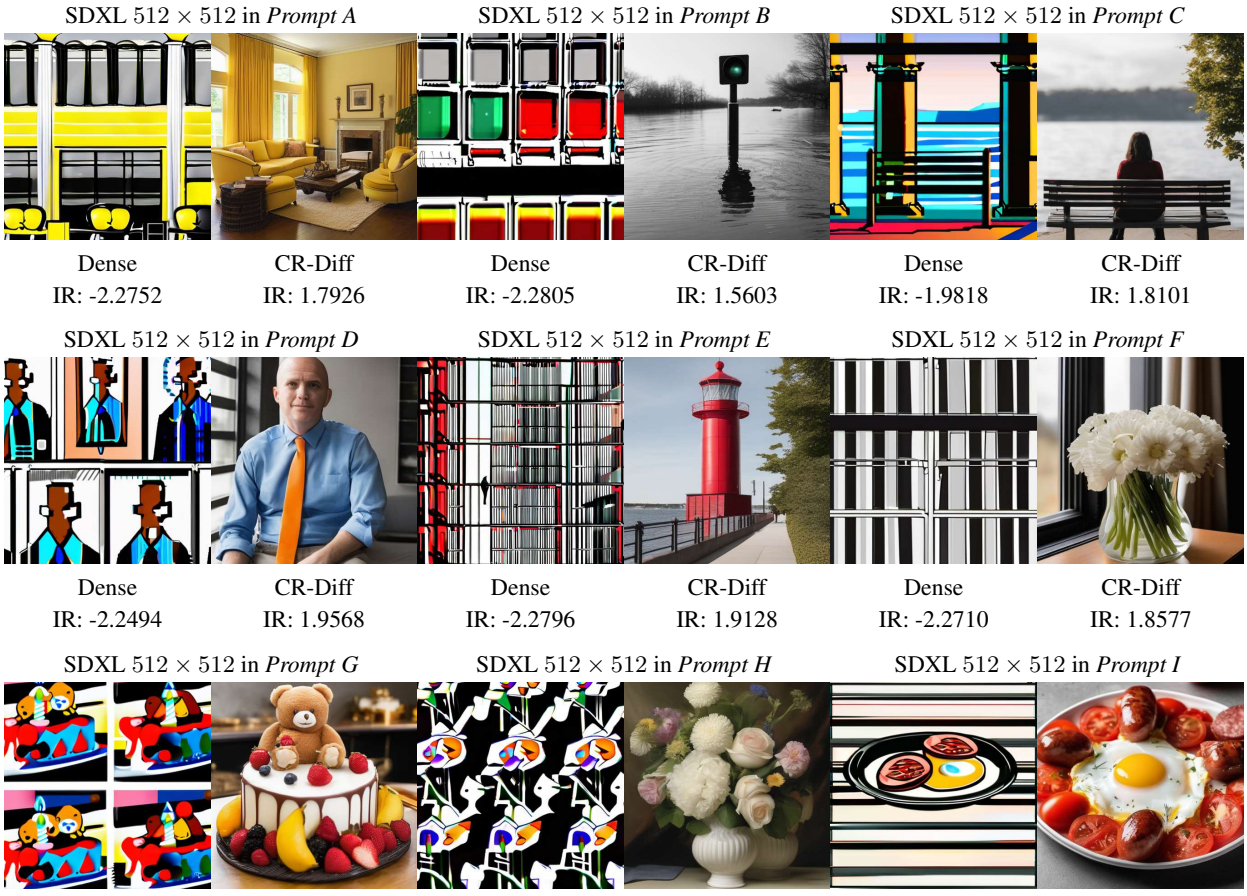
Data: $T_{init}, \alpha, N_{iter}, S_{seeds}, R_{max}$
Result: S_{best}
$(S_{curr}, E_{curr}) \leftarrow \text{BestSeed}(S_{seeds})$;
$S_{best} \leftarrow S_{curr}$; $E_{best} \leftarrow E_{curr}$;
$T \leftarrow T_{init}$; $C_{restart} \leftarrow 0$;
for $i = 1$ ***to*** N_{iter} **do**
  $S_{neighbor} \leftarrow \text{GenerateNeighbor}(S_{curr}, T)$;
  $E_{neighbor} \leftarrow \text{Evaluate}(S_{neighbor})$;
  if $E_{neighbor} < E_{curr}$ ***or*** $\exp(-(E_{neighbor} - E_{curr})/T) > \textit{rand}(0,1)$ **then**
    $S_{curr} \leftarrow S_{neighbor}$;
    $E_{curr} \leftarrow E_{neighbor}$;
  end
  if $E_{curr} < E_{best}$ **then**
    $S_{best} \leftarrow S_{curr}$;
    $E_{best} \leftarrow E_{curr}$;
  end
  $T \leftarrow \alpha \cdot T$;
  if T *is too small* **then**
    $T \leftarrow T_{init}$; $C_{restart} \leftarrow C_{restart} + 1$;
    if $C_{restart} > R_{max}$ **then**
      break
    end
  end
end
return S_{best};

Table 8. Evaluation of SDXL at the unseen higher resolution 1536×1536. The results show that the model remains stable at this scale and CR-Diff maintains comparable performance without altering semantic structure or perceptual characteristics.

Model	Resolution	FID ↓		CLIP ↑		ImageReward ↑		PickScore ↑		Aesthetic Score ↑	
		dense	ours	dense	ours	dense	ours	dense	ours	dense	ours
SDXL	1536×1536	46.563	40.380	0.315	0.312	0.300	0.208	21.675	21.399	5.952	5.855

Figure 9. Additional cross-resolution comparisons between SDXL [33] and its CR-Diff counterpart. CR-Diff consistently improves cross-resolution image quality compared to the dense model.

Figure 10. Additional cross-resolution comparisons between SDXL [33] and its CR-Diff counterpart. CR-Diff consistently improves cross-resolution image quality compared to the dense model.



Prompt A. A large living room with yellow curtains and couches

Prompt B. A traffic light emerged in a body of water.

Prompt C. A woman sits on a bench facing the water.

Prompt D. a man in a blue shirt and a orange tie

Prompt E. a tall red light house by the water and a walkway

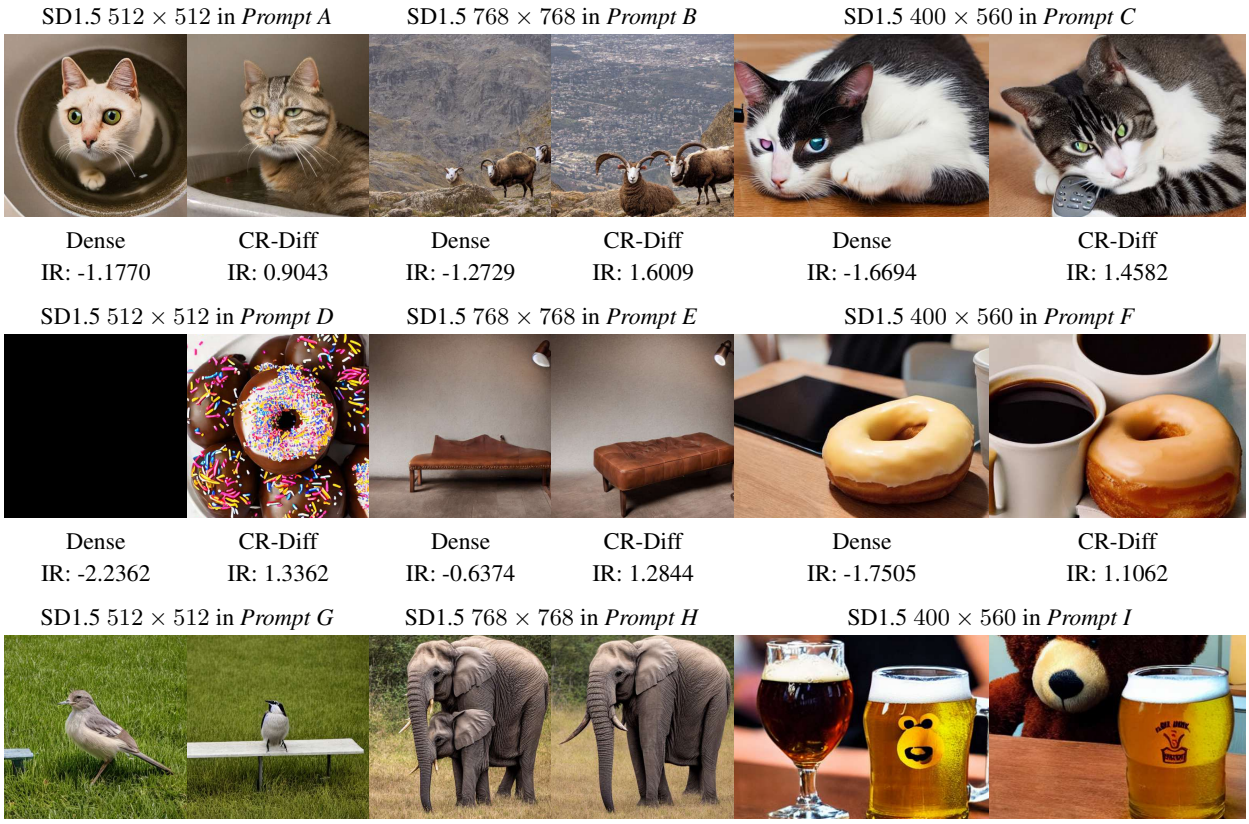
Prompt F. White flowers sit in a vase by the window.

Prompt G. birthday cake with fruit and teddy bears on it

Prompt H. White vase holding holding an assortment of flowers

Prompt I. a plate an egg tomato and sausage on it

Figure 11. Additional cross-resolution comparison on a subset of 5K prompts from the MS-COCO 2014 validation set [27]. CR-Diff shows consistent gains in both ImageReward and visual fidelity compared to the original SDXL. *Dense* denotes the original unpruned model. Each group corresponds to a specific prompt, and the ImageReward (IR) scores are shown below each image.



Prompt A. a close up of a cat in a bath room sink

Prompt B. Horned sheep high on a mountain with the city below

Prompt C. A cat laying down with its head on a remote control.

Prompt D. A white plate topped with donuts covered in chocolate and sprinkles.

Prompt E. A leather sitting bench is in a dimly lit room.

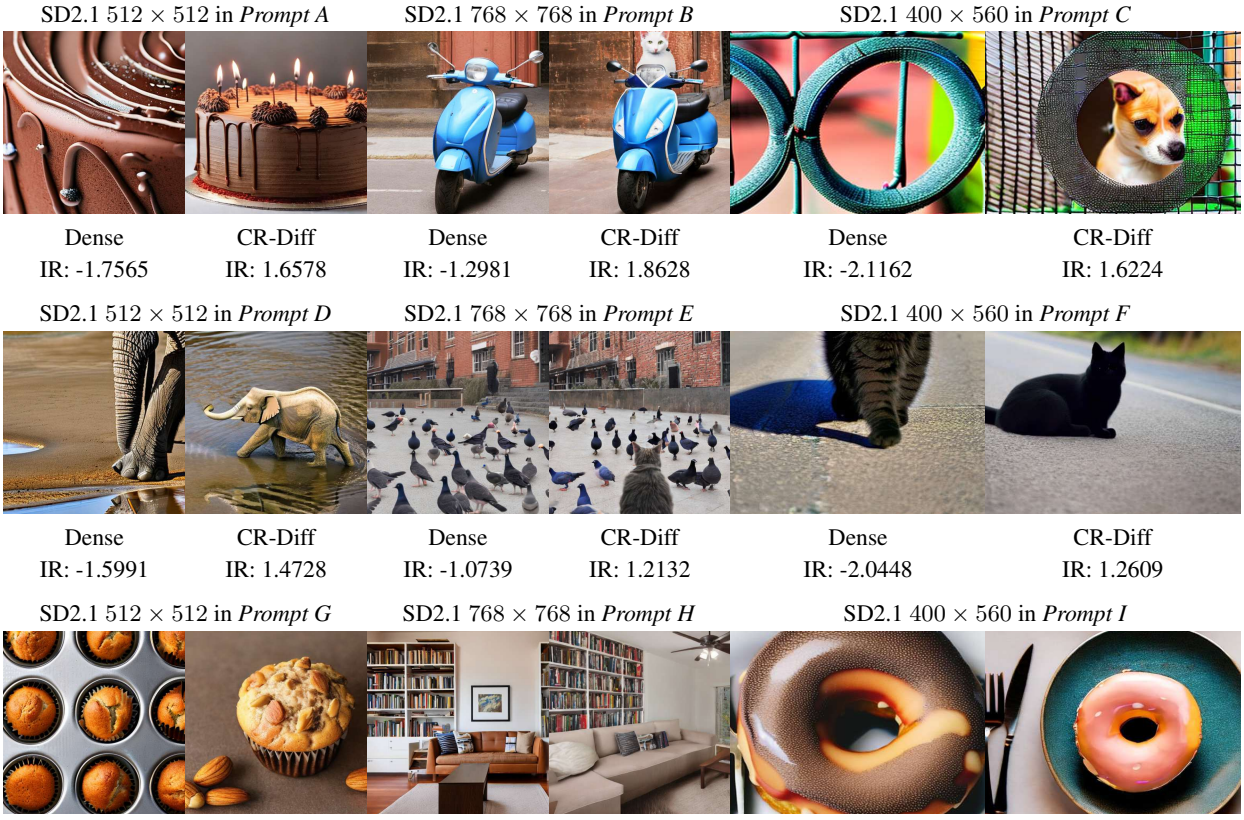
Prompt F. The large donut is next to a cup of coffee.

Prompt G. A bird is sitting on a bench in the grass.

Prompt H. A tusked elephant standing in an open field in the wild.

Prompt I. A teddy bear next to a glass of beer.

Figure 12. Additional cross-resolution comparison on a subset of 5K prompts from the MS-COCO 2014 validation set [27]. CR-Diff shows consistent gains in both ImageReward and visual fidelity compared to the original SD1.5. *Dense* denotes the original unpruned model. Each group corresponds to a specific prompt, and the ImageReward (IR) scores are shown below each image.



Prompt A. a chocolate frosting covered birthday cake with candles on it

Prompt B. A blue motor scooter with a cat sitting on it.

Prompt C. A dog looks out through a wire mesh hole in the fence.

Prompt D. A baby elephant walking through a shallow pool of flowing water

Prompt E. A cat is looking at a large group of pigeons.

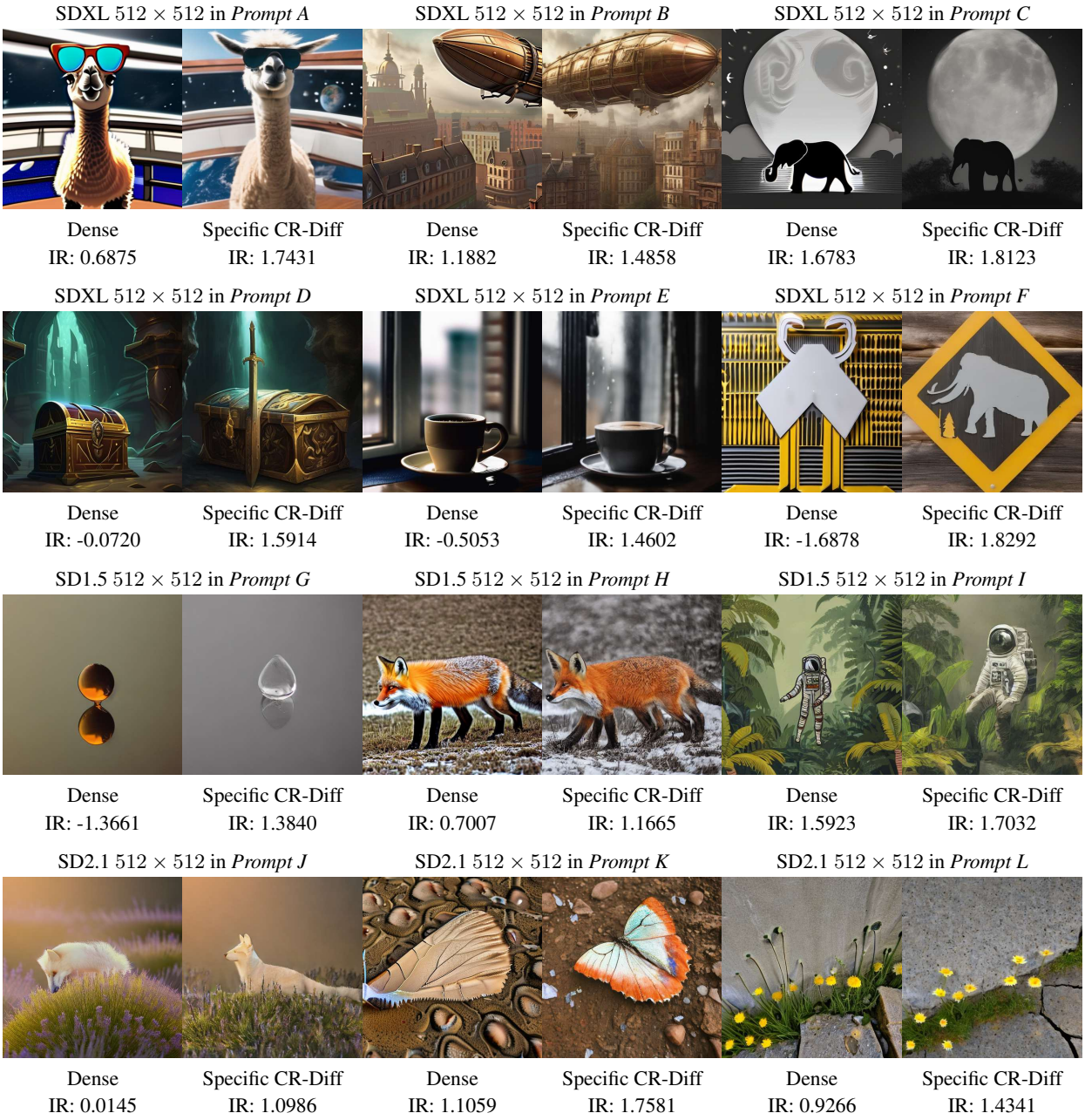
Prompt F. A black cat foolishly sitting in the middle of a road.

Prompt G. A muffin is adorned with a topping of nuts.

Prompt H. A room that has a couch, table, bookshelves and fan.

Prompt I. A donut on a plate with a fork and knife.

Figure 13. Additional cross-resolution comparison on a subset of 5K prompts from the MS-COCO 2014 validation set [27]. CR-Diff shows consistent gains in both ImageReward and visual fidelity compared to the original SD2.1. *Dense* denotes the original unpruned model. Each group corresponds to a specific prompt, and the ImageReward (IR) scores are shown below each image.



Prompt A. A photo of llama wearing sunglasses standing on the deck of a spaceship with the Earth in the background.

Prompt B. A steampunk airship docking above a Victorian city.

Prompt C. the silhouette of an elephant on the full moon

Prompt D. a painting of an ornate treasure chest with a broad sword propped up against it, glowing in a dark cave

Prompt E. A cup of coffee on a rainy window sill, cinematic lighting.

Prompt F. a yellow diamond-shaped sign with a wooly mammoth silhouette

Prompt G. A lone droplet on a reflective surface with shallow depth of field.

Prompt H. A fox in the snow, photograph.

Prompt I. Astronaut in a jungle, cold color palette, muted colors, detailed, 8k.

Prompt J. A white fox standing in a lavender field under gentle fog, rim light, dreamy haze, backlit sunlight, 50mm photography.

Prompt K. A butterfly wing fragment lying quietly on the earth.

Prompt L. A patch of wildflowers growing through a crack in concrete.

Figure 14. Visual comparison across two generation settings. *Dense* denotes the original unpruned model. *Specific CR-Diff* adjusts the pruning ratio based on each input prompt, enabling prompt-tailored optimization. Each group corresponds to a specific prompt, and the ImageReward (IR) scores are shown below each image. Both quantitative and qualitative results show that Specific CR-Diff improves semantic alignment and visual coherence for the given prompt.