Appendix

A. Additional Training Details

A.1. SCOP Dataset Construction

The optimal threshold values for the SCOP data engine were empirically determined to be $\{\tau_v, \tau_u, \tau_o, \tau_s\} = \{0.2, 2.0, 0.3, 0.5\}$ through grid search. We then applied these thresholds to process the COCO training split [43], resulting in the curated SCOP dataset. The initial Relationship Reasoning stage identifies 2,468,858 object pairs with their corresponding spatial relationships. The subsequent Spatial Constraints Enforcement stage systematically filters these pairs. This process applies a series of increasingly stringent criteria: Visual Significance eliminates 1,929,560 pairs, Semantic Distinction removes 169,973, Spatial Clarity filters out 119,457, Minimal Overlap excludes 148,376, and Size Balance removes a final 73,464 pairs. This cascade ultimately yields 28,028 clear and unambiguous object pairs.

During the training phase, the spatial relationships between object pairs are expressed with one of 8 distinct relationship tokens: `<left>`, `<above>`, `<right>`, `<below>`, `<left+above>`, `<right+above>`, `<left+below>`, and `<right+below>`. These relationships are determined based on the exact positions of the objects, with occasional random replacements using the `<and>` token to enhance robustness. For each pair, a square region containing both objects is automatically chosen, which is then randomly perturbed and expanded by up to 10% before being cropped to create the final training image paired with its corresponding text description.

A.2. Model Setups

Our experiments are conducted on four diffusion models, including three UNet-based diffusion models SD1.4[1], SD1.5[2], SD2.1[3], and the state-of-the-art MMDiT-based diffusion model FLUX.1-dev[4].

For the UNet-based models, we found that incorporating attention supervision as proposed by [71] significantly enhanced the convergence of the TENOR module. We therefore integrated this supervision across all UNet-based implementations. In contrast, for the FLUX.1 model, the standard denoising loss alone was sufficient to achieve optimal performance. Notably, our best results were achieved with the CoMPaSS-enhanced FLUX.1 model, using a rank-16 LoRA checkpoint that requires only ∼50MiB on-disk storage, making it highly efficient for practical applications.

Detailed training hyperparameters for all model configurations are provided in Tab. A7.

[1] https://huggingface.co/CompVis/stable-diffusion-v1-4

[2] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[3] https://huggingface.co/stabilityai/stable-diffusion-2-1

[4] https://huggingface.co/black-forest-labs/FLUX.1-dev

B. Runtime Performance of TENOR

The TENOR module represents the only potential source of additional computational overhead in our framework CoMPaSS, as it injects token ordering information into each text-image attention operation within the diffusion models. To quantify its impact, we conducted comprehensive benchmarking of inference latency across all model configurations:

Model	Latency @ 512×512	Overhead
SD1.4	1.17s ± 3.04ms	+0.85%
SD1.4 +TENOR	1.18s ± 7.24ms	
SD1.5	1.17s ± 2.50ms	+1.71%
SD1.5 +TENOR	1.19s ± 4.70ms	
SD2.1	1.13s ± 2.50ms	+4.42%
SD2.1 +TENOR	1.18s ± 4.70ms	
FLUX.1	17.3s ± 40.6ms	+2.89%
FLUX.1 +TENOR	17.8s ± 88.8ms	

Our measurements demonstrate that the TENOR module has minimal impact on runtime performance, introducing only negligible computational overhead. Even in the most demanding case of the FLUX.1-dev model, the additional time penalty amounts to just 2.89% of the total inference time, making it a highly practical enhancement for real-world applications.

C. Additional Results

C.1. Visualization of SCOP Data

We provide visualizations of data curated by SCOP in Fig. A8.

C.2. Additional Comparisons on VISOR

We present a comprehensive comparison of VISOR metrics against other state-of-the-art models in Tab. A10, demonstrating the superior performance of our approach across various evaluation criteria.

C.3. More Visual Comparisons

To provide a clear visualization of our approach's effectiveness, we present additional visual comparisons across eight spatial configurations in Figs. A9 to A12. For each prompt, we generate a total of 36 images using different model configurations, offering a comprehensive view of how our method consistently improves spatial understanding across various scenarios and model architectures.

Figure A8. **Example object pairs and their corresponding bounding boxes extracted by the SCOP data engine.** Each pair satisfies our spatial constraints for Visual Significance, Semantic Distinction, Spatial Clarity, Minimal Overlap, and Size Balance, ensuring unambiguous spatial relationships.

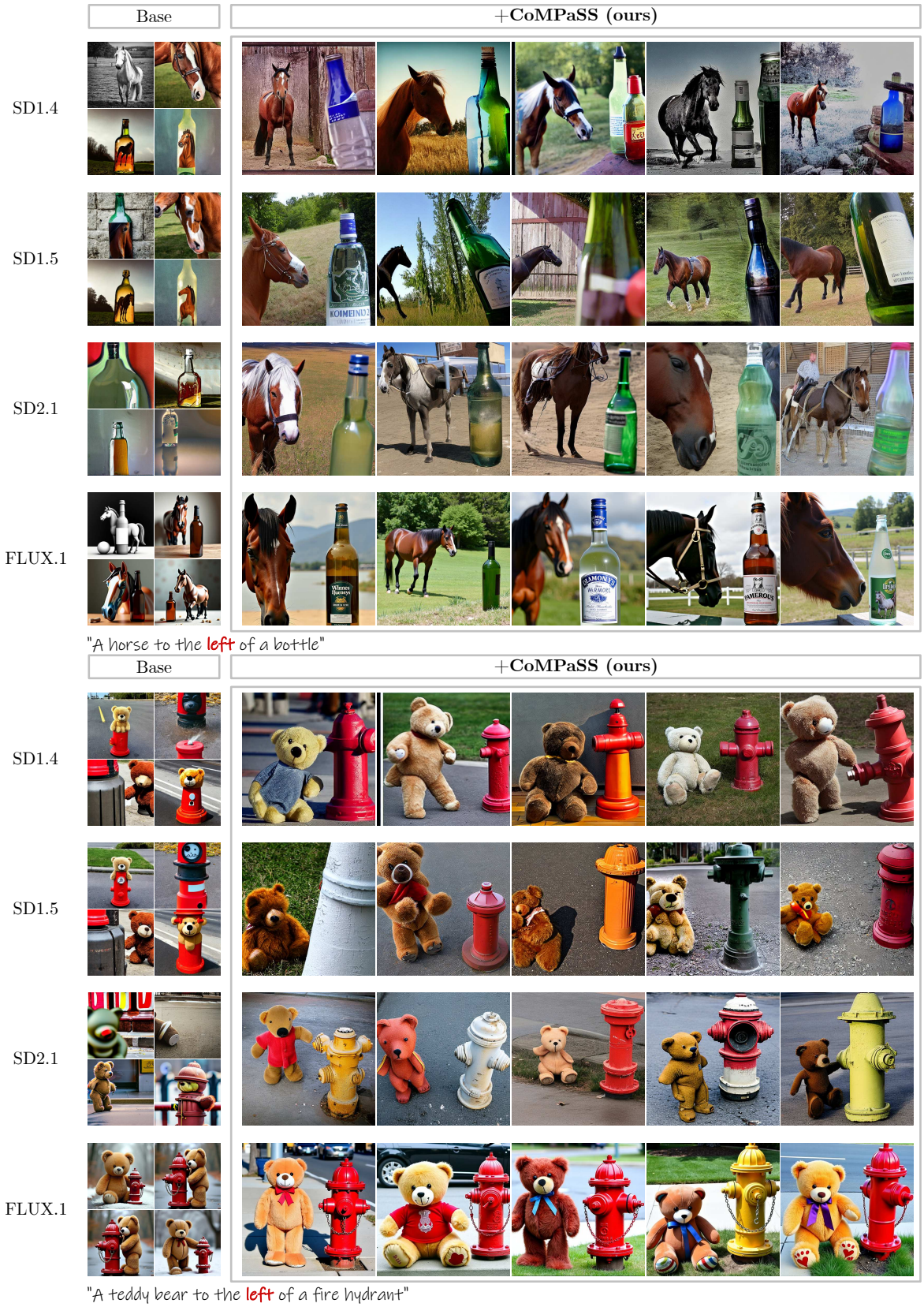

Figure A9. **Additional results demonstrating spatial relationship "`left`".** Our method consistently improves spatial accuracy over the baseline models.

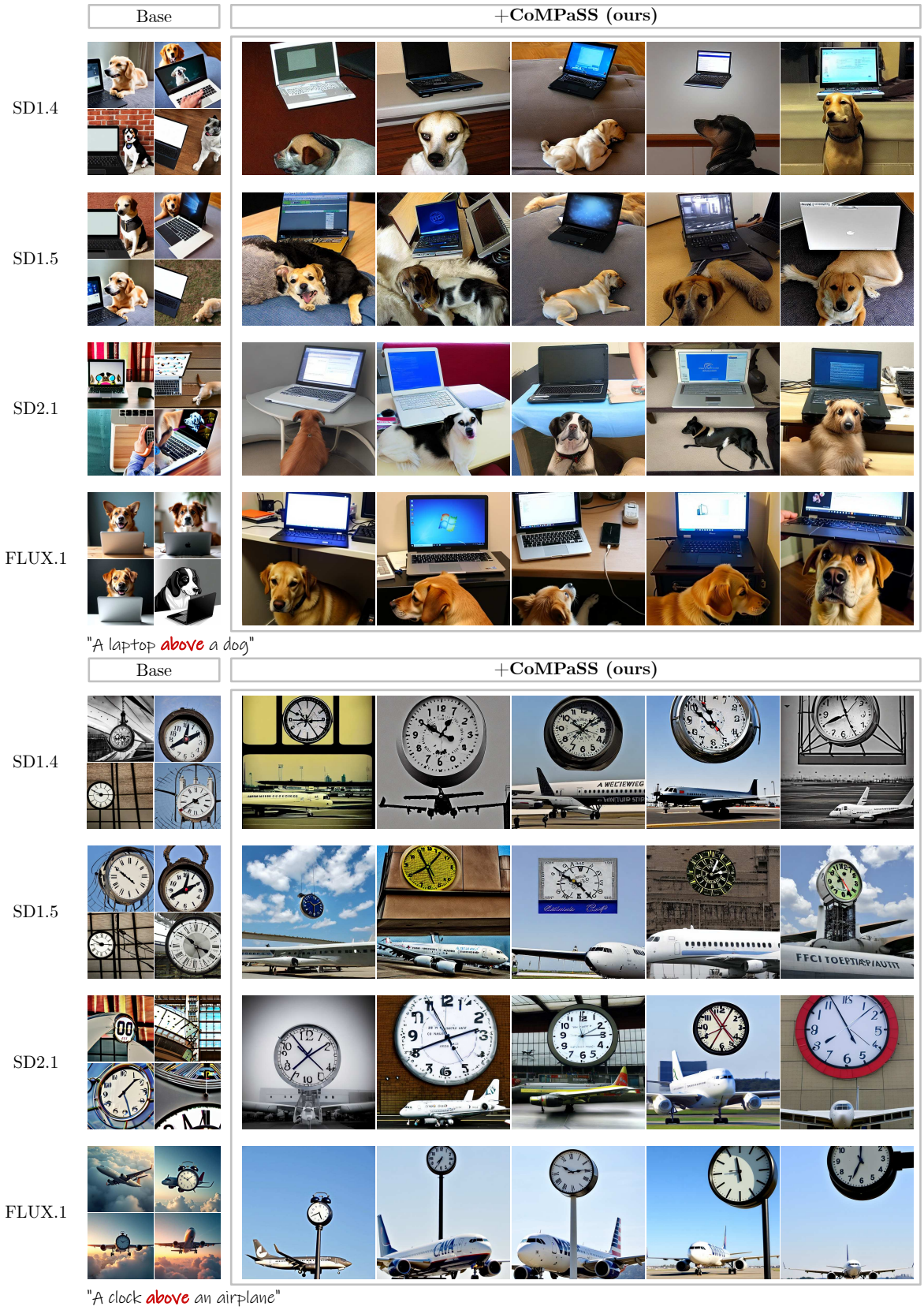

Figure A10. **Additional results demonstrating spatial relationship "above".** Our method consistently improves spatial accuracy over the baseline models.

Figure A11. **Additional results demonstrating spatial relationship "`right`".** Our method consistently improves spatial accuracy over the baseline models.

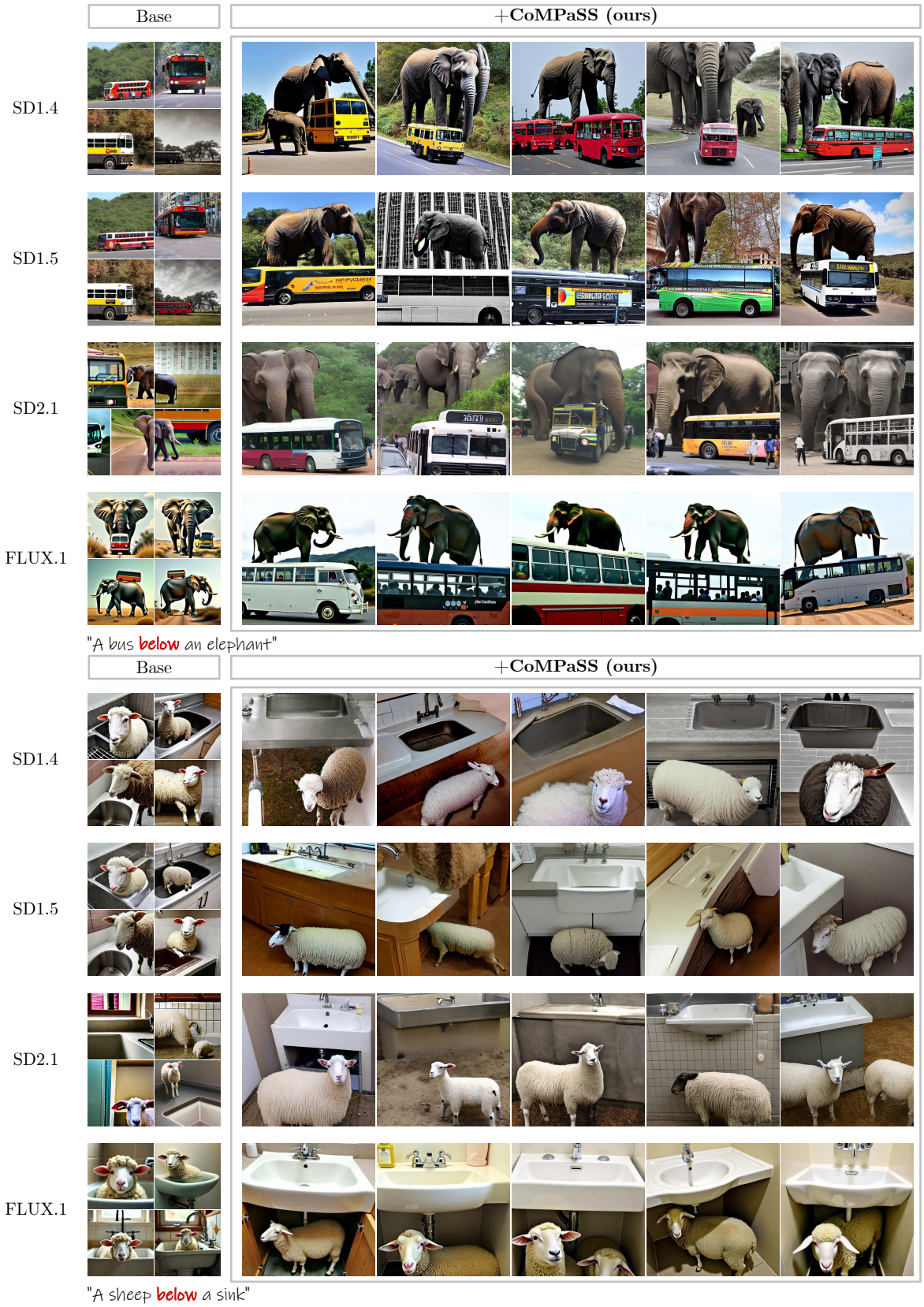

Figure A12. **Additional results demonstrating spatial relationship "`below`".** Our method consistently improves spatial accuracy over the baseline models.

Table A7. **Hyperparameters used during training.**

Hyperparameter	**SD1.4**	**SD1.5**	**SD2.1**	**FLUX.1**
AdamW Learning Rate (LR)	5e-6	5e-6	5e-6	1e-4
AdamW β_1	0.9	0.9	0.9	0.9
AdamW β_2	0.999	0.999	0.999	0.999
AdamW ϵ	1e-8	1e-8	1e-8	1e-8
AdamW Weight Decay	1e-2	1e-2	1e-2	1e-2
LR scheduler	Constant	Constant	Constant	Constant
LR warmup steps	0	0	0	20
Training Steps	24,000	24,000	80,000	24,000
Local Batch Size	1	1	1	1
Gradient Accumulation	2	2	2	2
Training GPUs	2×L40S	2×L40S	2×L40S	2×L40S
Training Resolution	512×512	512×512	512×512	512×512
Trained Parameters	All parameters of diffusion UNet			LoRA (rank=16) on all DoubleStreamBlocks
Prompt Dropout Probability	10%	10%	10%	10%

Table A8. **Multi-object spatial relationship evaluation.** Despite being trained only on object pairs, CoMPaSS improves the model's spatial accuracy on prompts involving three objects.

Model	% Correct (any)	% Correct (all)
SD2.1	9.1	3.3
SD2.1 **+CoMPaSS**	**32.3**	**20.8**
FLUX.1	30.12	13.1
FLUX.1 **+CoMPaSS**	**52.44**	**31.2**

C.4. Evaluation on Three-Object Compositions

To further test the generalization capabilities of CoMPaSS, we constructed a more challenging benchmark focused on three-object spatial configurations. This benchmark contains 512 prompts, such as "`a clock above a broccoli, with a car to the left of the clock`", and is evaluated using an extension of the GenEval methodology [25]. Despite not being fine-tuned on such complex prompts, CoMPaSS demonstrates a significant improvement in accuracy, as detailed in Tab. A8. This result builds upon the strong two-object performance shown in the main paper (Tab. 2).

C.5. Detailed Benchmark Breakdown

While the main paper reported overall scores for standard benchmarks (Tab. 2), here we provide a more granular breakdown. Table A11 presents the full, per-task results for T2I-CompBench, GenEval [25], and DPG-Bench [33], offering a comprehensive performance overview.

C.6. Ablation Studies on other Models

While Tab. 6 in the main paper presented ablation studies on SD1.5 and FLUX.1, here we extend our analysis to the other two UNet-based diffusion models SD1.4 and SD2.1 in Tab. A9. The results consistently demonstrate that both components of CoMPaSS contribute to improved spatial understanding across different model architectures.

Table A9. **Ablation study on the components of CoMPaSS.** (i) original models; (ii) trained with the SCOP dataset described in Sec. 3.1 of the main paper; (iii) our full method. T2I-CompBench Spatial (T. Spatial) and GenEval Position (G. Pos) scores are reported.

Setting	Model	Components		T. Spatial	G. Pos.
		SCOP	TENOR		
(i)	SD1.4			0.12	0.03
(ii)	SD1.4	✓		0.29	0.36
(iii)	SD1.4	✓	✓	**0.34**	**0.46**
(i)	SD2.1			0.13	0.07
(ii)	SD2.1	✓		0.30	0.36
(iii)	SD2.1	✓	✓	**0.32**	**0.51**

Table A10. **Comparison to state-of-the-art models on the VISOR [26] benchmark.** OA stands for "object accuracy", which measures the rate at which all prompted objects appear in the generated image.

Method	uncond	cond	1	2	3	4	OA
GLIDE [49]	1.98	59.06	6.72	1.02	0.17	0.03	3.36
DALLE-mini [17]	16.17	59.67	38.31	17.50	6.89	1.96	27.10
CogView2 [19]	12.17	65.89	33.47	11.43	3.22	0.57	18.47
Structured Diffusion [22]	17.87	62.36	44.70	18.73	6.57	1.46	28.65
DALLE-2 [57]	37.89	59.27	73.59	47.23	23.26	7.49	63.93
SD1.4	18.81	62.98	46.60	20.11	6.89	1.63	29.86
SD1.5	17.58	61.08	43.65	18.62	6.49	1.57	28.79
SD2.1	30.25	63.24	64.42	35.74	16.13	4.70	47.83
SD2.1 +SPRIGHT [6]	43.23	71.24	71.78	51.88	33.09	16.15	60.68
FLUX.1	37.96	66.81	64.00	44.18	28.66	14.98	56.95
SD1.4 +**CoMPaSS**	57.41	87.58	83.23	67.53	49.99	28.91	65.56
SD1.5 +**CoMPaSS**	61.46	**93.43**	86.55	72.13	54.64	32.54	65.78
SD2.1 +**CoMPaSS**	62.06	90.96	85.02	71.29	56.03	35.90	68.23
FLUX.1 +**CoMPaSS**	**75.17**	93.22	**91.73**	**83.31**	**72.21**	**53.41**	**78.64**

Table A11. **Evaluation results of general generation capabilities across a wide range of tasks on GenEval [25], T2I-CompBench [35], and DPG-Bench [33].** While designed to target spatial performance, CoMPaSS also improves overall (Ovr.) alignment scores across most tasks.

Method	T2I-CompBench					GenEval							DPG-Bench					
	Spat.	Col.	Shp.	Tex.	N.Sp.	**Pos.**	S.O.	T.O.	Count	Col.	Attr.	**Ovr.**	**Rel.**	Glb.	Ent.	Attr.	Other	**Ovr.**
SD1.4	0.12	0.38	0.36	0.42	0.31	0.03	0.98	0.41	0.34	0.74	0.06	0.43	81.04	78.12	72.23	72.56	59.60	62.02
SD1.4 +**CoMPaSS**	**0.34**	0.49	0.43	0.53	0.31	**0.46**	0.99	0.68	0.34	0.73	0.17	**0.56**	**83.21**	79.33	75.33	72.09	68.00	**66.07**
SD1.5	0.08	0.38	0.37	0.42	0.31	0.04	0.96	0.38	0.36	0.75	0.06	0.42	73.49	74.63	74.23	75.39	67.81	63.18
SD1.5 +**CoMPaSS**	**0.35**	0.50	0.43	0.52	0.31	**0.54**	0.99	0.69	0.34	0.72	0.15	**0.57**	**84.10**	82.67	75.20	73.58	60.80	**65.81**
SD2.1	0.13	0.51	0.42	0.49	0.31	0.07	0.98	0.51	0.44	0.85	0.17	0.50	83.95	81.16	74.47	75.29	53.60	65.47
SD2.1 +SPRIGHT [6]	0.21	-	-	-	-	0.11	0.99	0.59	0.49	0.85	0.15	0.51	-	-	-	-	-	-
SD2.1 +**CoMPaSS**	**0.32**	0.55	0.43	0.54	0.30	**0.51**	0.99	0.69	0.20	0.71	0.15	**0.54**	**86.54**	79.94	78.89	75.39	62.80	**69.48**
FLUX.1	0.18	0.69	0.48	0.63	0.31	0.26	0.92	0.77	0.71	0.66	0.27	0.60	92.30	80.55	87.74	85.55	78.40	80.63
FLUX.1 +**CoMPaSS**	**0.30**	0.83	0.59	0.71	0.32	**0.60**	0.99	0.87	0.71	0.80	0.76	**0.76**	**94.12**	82.98	90.53	88.30	82.80	**84.42**