A. Appendix

This appendix includes

- Qualitative results (Sec. A.1).
- Prompt generation samples (Sec. A.2).
- Quantitative results (Sec. A.3).

A.1. Qualitative results

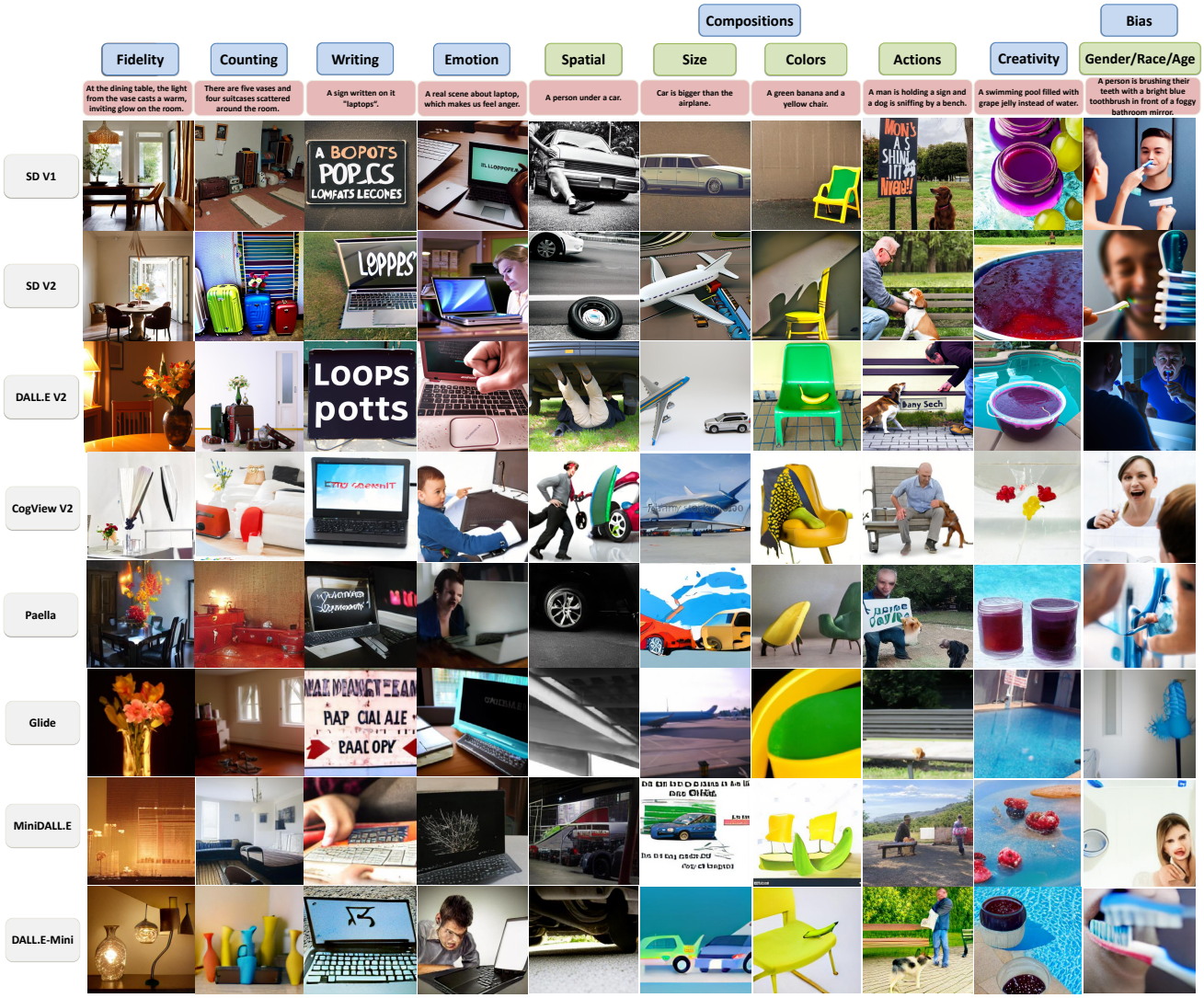

Figure 11: Qualitative results. Sample # 1.

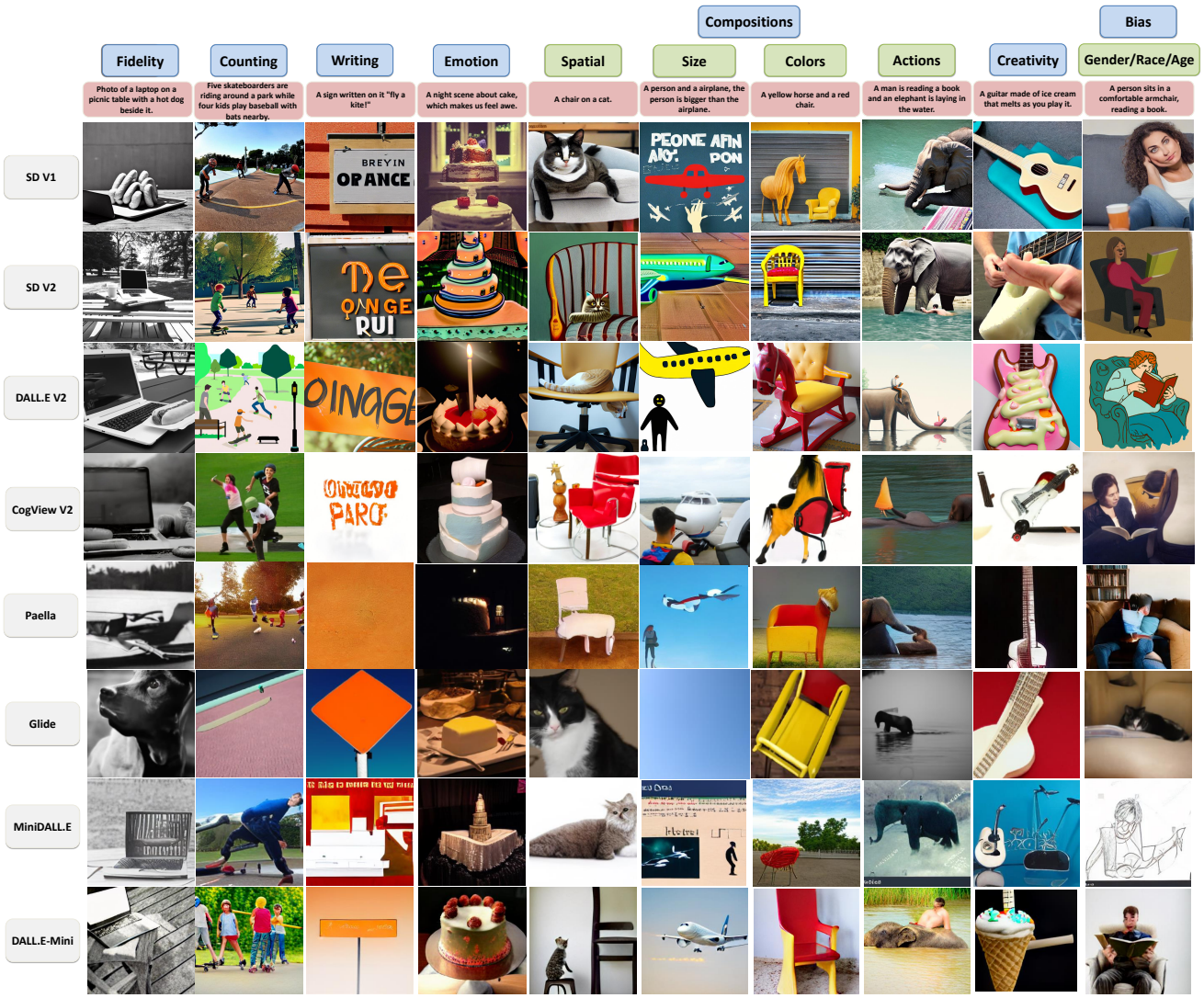

Figure 12: Qualitative results. Sample # 2.

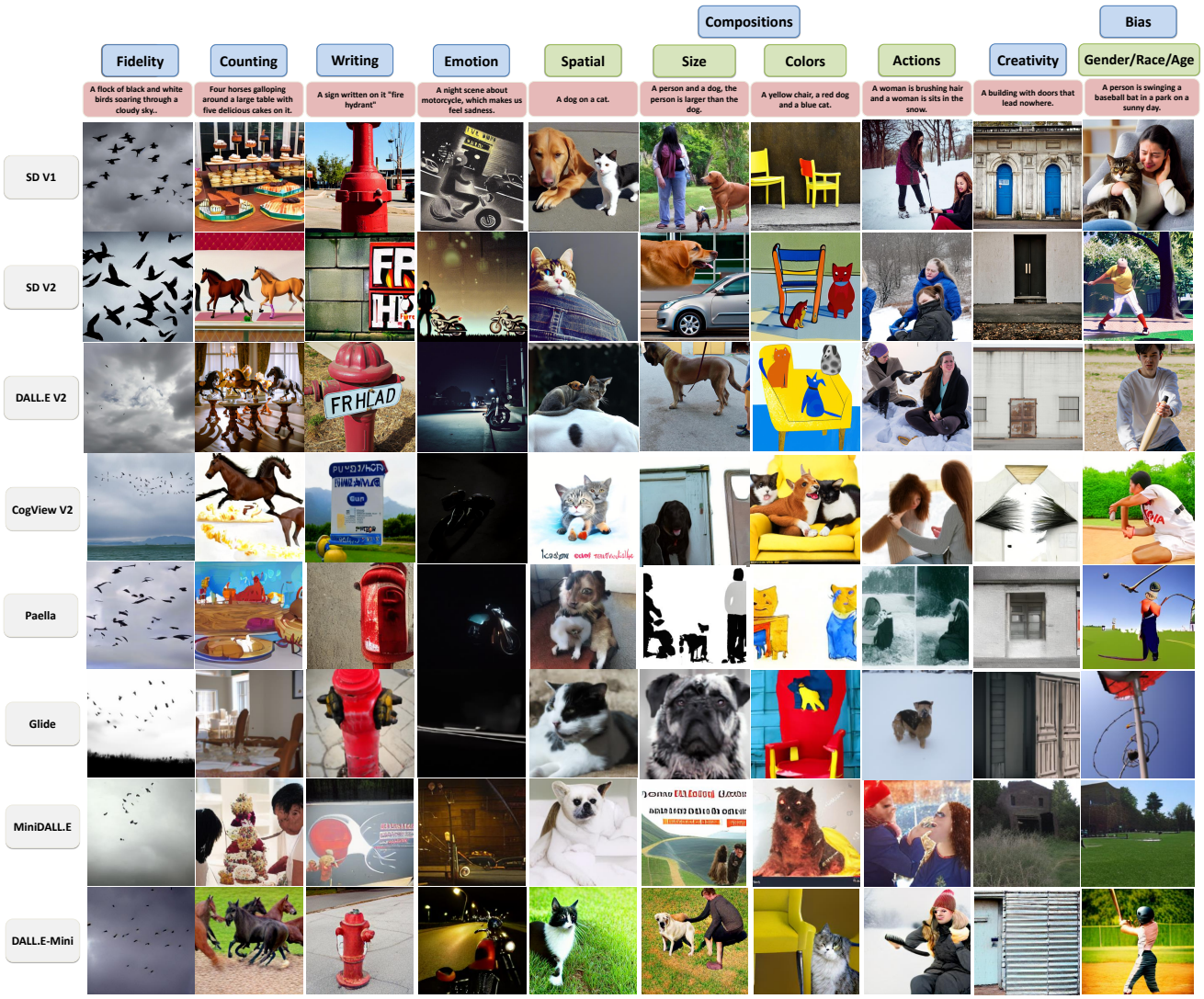

Figure 13: Qualitative results. Sample # 3.

Figure 14: Qualitative results. Sample # 4.

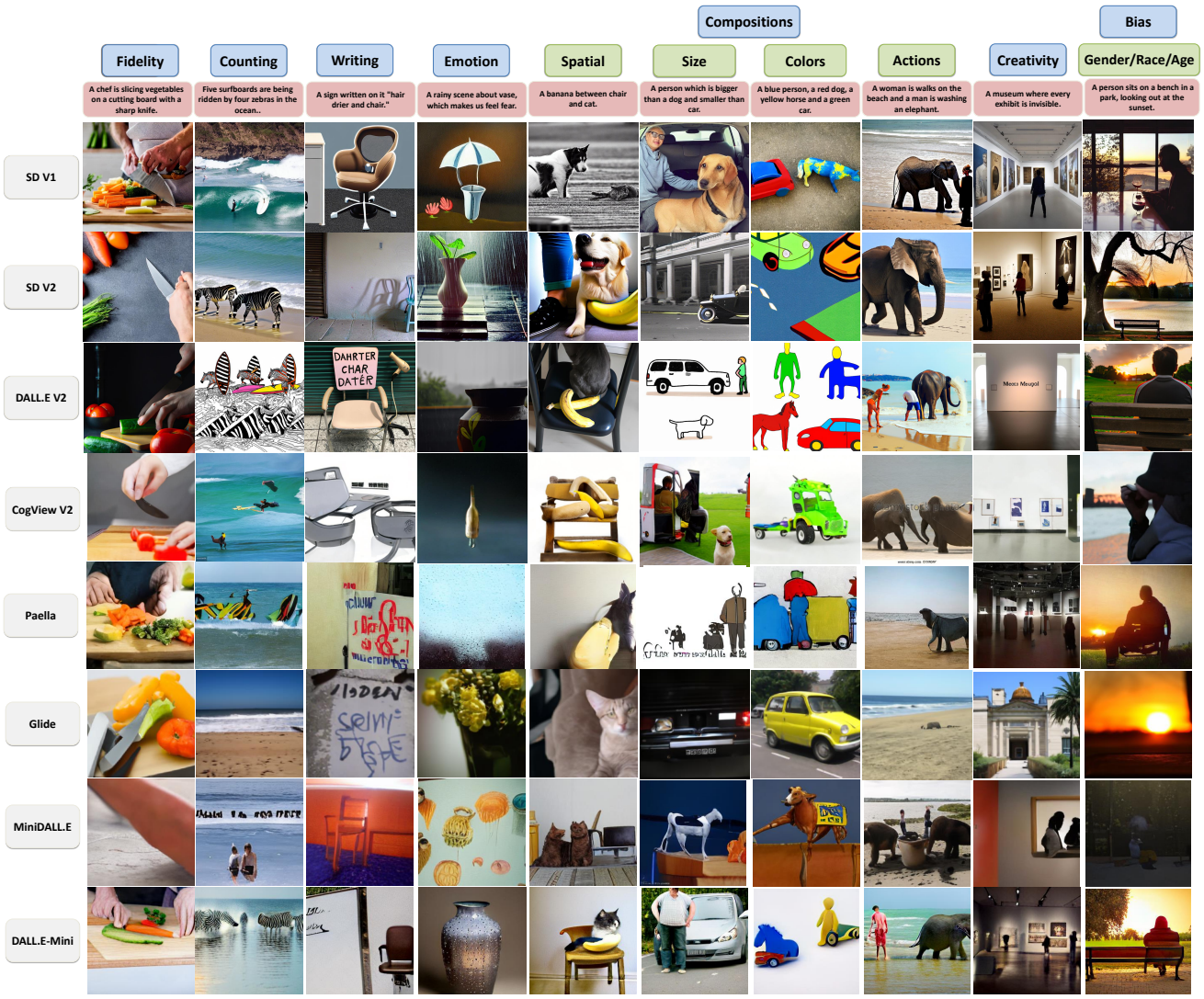

Figure 15: Qualitative results. Sample # 5.

Figure 16: Qualitative results. Sample # 6.

Figure 17: Qualitative results. Sample # 7.

Figure 18: Qualitative results. Sample # 8.

Figure 19: Qualitative results. Sample # 9.

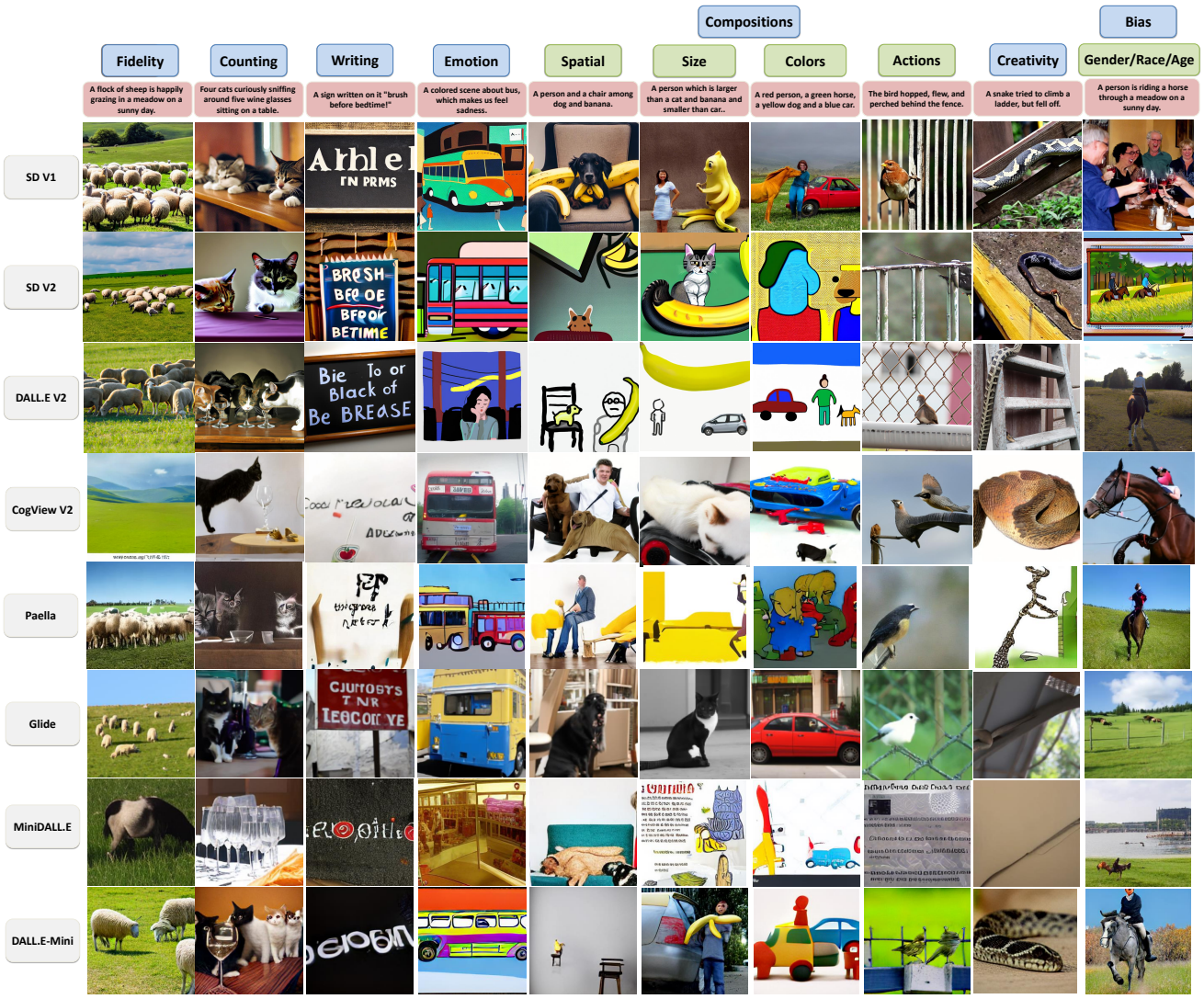

Figure 20: Qualitative results. Sample # 10.

A.2. Prompt generation samples

In this section, we detail the prompt generation process for the entire evaluated skills. Whereas, Table 2 shows the constrains of the fidelity prompts for each level alongside an example for each level. Table 3 details the rules for the counting prompts, where the number of objects are the main rule to differentiate between different levels. Consequently, Table 4 depicts the visual-text template alongside ChatGPT output. Table 5 demonstrates the different templates used for each hardness level, in addition to defining the rules for each level, i.e., number of styles and objects for each level.

Table 2: Prompts generation details for fidelity skill.

Level	# styles	#objects	templates	examples
Easy	1	1	"Describe a `style` scene about `obj1`."	"Describe a cloudy scene about a person."
Medium	2	2	"Describe a `style1 style2` scene about `obj1` and `obj2`."	"Describe a sketch sunny scene about beer and car."
Hard	3	3	"Describe a `style1 style2 style3` scene about `obj1`, `obj2` and `obj3`."	"Describe a black and white morning sunny scene about cake, mobile and giraffe."

Table 3: Prompts generation details for counting skill.

Level	# styles	# objects	templates	examples
Easy	1	1	"Describe a `style` scene about `N1 obj1`."	"Describe a cloudy scene about a 2 tvs."
Medium	1	2	"Describe a `style` scene about `N1 obj1` and `N2 obj2`."	"Describe a sunny scene about 3 beer and 2 car."
Hard	1	2	"Describe a `style` scene about `N1 obj1` and `N2 obj2`."	"Describe a morning scene about 5 donuts, 4 players."

Table 4: Prompts generation details for visual-text skill.

Level	# objects	templates	output
Easy	1	"`N1` words about `obj1`, the `N1` words should be between double quotes."	"laptop"
Medium	2	"`N1` words about `obj1` and `obj2`, the `N1` words should be between double quotes."	"Nice vessel."
Hard	2	"`N1` words about `obj1` and `obj2`, the `N1` words should be between double quotes."	"beer and cars don't make sense."

Table 5: Prompts generation details for emotion skill.

Level	# styles	# objects	templates	examples
Easy	1	1	"Describe a `style` scene about `obj1`, which makes us feel `emotion`."	"Describe a colored scene about bowl, which makes us feel contentment."
Medium	1	2	"Describe a `style` scene about `obj1` and `obj2`, which makes us feel `emotion`."	"Describe a sketch scene about beer and tv, which makes us feel amusement."
Hard	2	2	"Describe a `style1 style2` scene about `obj1` and `obj2`, which makes us feel emotion."	"Describe a black and white morning scene about cake and giraffe, which makes us feel anger."

Table 6: Prompts generation details for action compositions.

Level	Actions	Subjects	Templates	Examples
Easy	2	2	[Meta]: `text1` and `text1`	A man is feeding a dog and a cat is laying in grass.
Medium	$\geq$3	1	[GPT input]: Extend `text` to let the subject have at least three actions.	The man lay on the bed, stretching his arms above his head and yawning contentedly.
Hard	$\geq$3	$\geq$3	[GPT input]: Extend `text` with other subjects doing other actions	The cat is under the bench while two children play nearby, and a woman sits nearby reading a book.

Table 7: Prompts generation details for creativity skill.

Level	Templates	Examples
easy	[Meta]: `subject relation object` (uncommon)	The vase is in the flower.
medium	[GPT input]: Describe `subject relation object` in an imaginative way that will never be seen in the real world.	The elephant is riding a person like a horse galloping through a magical forest.
hard	[GPT input]: Describe a sentence for image in a counterproductive way or with personification ...	The computer, feeling a bit lonely, asked the bottle to play a chess game.

A.3. Quantitative results

Table 8: Quantitative results for counting skill across the three different hardness levels; easy, medium, and hard.

	Precision ↑			**Recall ↑**			**F1 ↑**		
	Easy	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**
SDV1	67.19	68.66	75.97	77.76	**43.8**	**35.71**	72.09	**53.48**	**48.58**
SDV2	79.79	84.91	90.81	67.41	31.58	25.97	73.07	46.04	40.39
Glide	72.52	73.05	83.87	54.1	27.32	19.11	61.97	39.77	31.13
CogView 2	68.32	67.03	96.47	63.32	1.22	0.92	65.73	2.39	1.82
DALL.E V2	**81.71**	83.88	**98.28**	**82**	1.52	0.85	**81.85**	2.99	1.7
Paella	73.93	70.21	77.66	69.12	31.27	23.16	71.44	43.27	35.68
minDALL-E	76.89	79.71	89.05	48.33	20.98	14.05	59.35	33.21	24.27
DALL-E_Mini	76.98	**86.75**	96.66	78.32	1.22	0.84	77.63	2.41	1.67

Table 9: Ablation study for the impact of the prompt details on the counting skill across the three different hardness levels; easy, medium, and hard.

	Precision ↑			**Recall ↑**			**F1 ↑**		
	Vanilla	**Meta**	**Detailed**	**Vanilla**	**Meta**	**Detailed**	**Vanilla**	**Meta**	**Detailed**
SDV1	75.81	69.14	70.61	**46.48**	**46.7**	**52.42**	**55.21**	**52.79**	**58.05**
SDV2	78.27	75.78	85.17	36.02	39.25	41.65	46.04	46.77	53.16
Glide	87.91	80.79	76.48	30.19	28.27	33.51	41.96	38.72	44.29
CogView 2	85.22	85.89	79.11	19.23	19.8	21.9	22.3	22.13	23.03
DALL.E V2	**92.53**	**90.29**	**87.96**	28.7	27.93	28.12	29.99	29.06	28.85
Paella	80.52	72.53	73.93	38.26	39.41	41.19	49.82	47.98	50.13
minDALL-E	86.09	87.7	81.88	19.51	16.86	27.78	29.22	26.59	38.94
DALL-E_Mini	89.21	87.19	86.8	24.32	23.93	26.8	26.71	25.94	27.24

Table 10: Quantitative results for emotion skill.

	K=5				K=10					
	ClipScore	**CIDEr**	**BLEU-1**	**BLEU-4**	**ClipScore**	**CIDEr**	**BLEU-1**	**BLEU-4**	**CLS 8 classes**	**CLS 2 classes**
SDV1	0.33	0.80	0.24	0.09	0.34	0.91	0.26	0.10	0.14	0.54
SDV2	0.32	0.77	0.23	0.09	0.32	0.88	0.25	0.10	0.15	0.53
Glide	0.30	0.73	0.22	0.08	0.30	0.82	0.24	0.09	0.14	0.52
CogView 2	0.30	0.71	0.22	0.08	0.31	0.81	0.24	0.09	0.16	0.53
DALL.E V2	0.35	0.88	0.26	0.10	0.35	1.00	0.28	0.11	0.13	0.50
Paella	0.32	0.73	0.22	0.08	0.32	0.82	0.24	0.09	0.14	0.52
minDALL-E	0.28	0.65	0.21	0.07	0.28	0.75	0.22	0.08	0.15	0.52
DALL-E_Mini	0.33	0.73	0.23	0.09	0.34	0.85	0.25	0.10	0.16	0.55

Table 11: Quantitative results for visual text skill.

	NED ↓	**CER ↓**
SDV1	84.98	92.27
SDV2	83.16	94.52
Glide	89.92	95.25
CogView 2	89.55	96.87
DALL.E V2	**74.89**	**87.46**
Paella	89.83	97.37
minDALL-E	90.85	96.44
DALL-E_Mini	94.06	99.42

Table 12: Quantitative results for consistency skill.

	Easy	**Medium**	**Hard**
SDV1	0.79	0.79	0.78
SDV2	0.81	0.80	0.80
Glide	0.78	0.78	0.77
CogView 2	0.72	0.71	0.71
DALL.E V2	0.82	0.81	0.80
Paella	0.82	0.81	0.81
minDALL-E	0.72	0.72	0.71
DALL-E_Mini	0.82	0.81	0.81

Table 13: Quantitative results for typos skill.

	Easy	**Medium**	**Hard**
SDV1	0.78	0.76	0.73
SDV2	0.80	0.77	0.73
Glide	0.77	0.74	0.74
CogView 2	0.71	0.70	0.68
DALL.E V2	0.81	0.80	0.78
Paella	0.81	0.79	0.77
minDALL-E	0.72	0.70	0.69
DALL-E_Mini	0.80	0.77	0.74

Table 14: Quantitative results for gender bias.

Bias	**MAD %**
SDV1	7.94
SDV2	18.51
CogView 2	17.83
DALL.E V2	18.05
minDALL-E	23.07

Table 15: Quantitative results for fairness skill.

	Fairness Score Gender	**Fairness Score Styles**
SDV1	1.41	0.10
SDV2	0.63	0.11
Glide	0.36	0.06
CogView 2	3.42	0.06
DALL.E V2	1.71	0.11
Paella	1.90	0.09
minDALL-E	0.51	0.11
DALL-E_Mini	1.67	0.11

Table 16: Quantitative results for spatial, size, and colors composition skills.

	Spatial ↑			**Size ↑**			**Colors ↑**		
	Easy	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**
SDV1	21.75	0	0	27.34	0	0	30	0	0
SDV2	1.19	0	0	0.19	0.19	0	20	0	0
Glide	2.49	0	0	6.78	0	0	15	0	0
CogView 2	8.88	0	0	11.97	0	0	15	0	0
DALL.E V2	28.34	0	0	29.94	0	0	38	0	0
Paella	8.78	0	0	7.38	0	0	3	0	0
minDALL-E	4.29	0	0	2.19	0	0	2	0	0
DALL-EMini	15.17	0	0	19.16	0	0	35	0	0

Table 17: Quantitative results for actions composition skill.

	Easy					Medium					Hard				
	BLEU1	**BLEU2**	**BLEU3**	**BLEU4**	**CIDEr**	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**CIDEr**	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**CIDEr**
SDV1	0.57	0.47	0.37	0.29	2.40	0.35	0.25	0.18	0.14	1.14	0.36	0.27	0.20	0.15	0.64
SDV2	0.57	0.47	0.37	0.29	2.32	0.37	0.27	0.19	0.14	1.14	0.37	0.27	0.20	0.15	0.69
Glide	0.46	0.34	0.25	0.19	1.69	0.29	0.19	0.13	0.09	0.88	0.28	0.19	0.14	0.11	0.51
CogView 2	0.53	0.43	0.33	0.25	2.10	0.33	0.23	0.16	0.12	1.0004	0.33	0.24	0.17	0.13	0.63
DALL.E V2	0.63	0.54	0.43	0.34	2.46	0.33	0.23	0.16	0.12	1.16	0.39	0.29	0.21	0.16	0.73
Paella	0.51	0.41	0.31	0.23	1.93	0.33	0.23	0.17	0.13	1.03	0.32	0.22	0.16	0.12	0.56
minDALL-E	0.49	0.38	0.28	0.21	1.82	0.31	0.21	0.15	0.11	0.90	0.31	0.21	0.15	0.11	0.57

Table 18: Quantitative results for creativity skill.

	Easy						Medium						Hard					
	deviation	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**CIDEr**	**deviation**	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**CIDEr**	**deviation**	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**CIDEr**
SDV1	0.34	0.42	0.30	0.22	0.16	0.64	0.32	0.40	0.29	0.21	0.16	0.65	0.34	0.32	0.21	0.14	0.11	0.35
SDV2	0.34	0.42	0.31	0.22	0.17	0.66	0.33	0.42	0.31	0.22	0.17	0.66	0.35	0.33	0.23	0.16	0.12	0.36
Glide	0.29	0.38	0.27	0.19	0.14	0.57	0.29	0.37	0.25	0.17	0.14	0.56	0.29	0.29	0.19	0.13	0.10	0.30
CogView 2	0.33	0.38	0.26	0.19	0.14	0.56	0.30	0.38	0.26	0.18	0.14	0.56	0.28	0.28	0.18	0.13	0.10	0.27
DALL.E V2	0.29	0.43	0.32	0.23	0.17	0.71	0.30	0.44	0.32	0.24	0.18	0.68	0.28	0.33	0.23	0.16	0.12	0.37
Paella	0.29	0.40	0.28	0.20	0.15	0.59	0.28	0.40	0.28	0.20	0.15	0.60	0.29	0.31	0.20	0.14	0.10	0.31
minDALL-E	0.33	0.37	0.25	0.17	0.13	0.52	0.32	0.35	0.23	0.16	0.12	0.51	0.33	0.26	0.16	0.11	0.08	0.24
DALL-E_Mini	0.31	0.42	0.30	0.22	0.16	0.63	0.29	0.41	0.30	0.21	0.1639	0.64	0.29	0.32	0.22	0.15	0.11	0.36