Evaluating Text-to-Image Synthesis with a Conditional Fréchet Distance

Supplementary Material

A. Image and Text Backbone Models

In our experiments, we used 46 different image backbone models and 43 different text backbone models. For vision models, we employed those trained with self-supervised learning, including ViT trained on ImageNet-1k [9], ViT trained on ImageNet-21k [35], MAE [16], DINOv2 [26], MoCOv3 [2], and I-JEPA trained on both ImageNet-1k and ImageNet-22k [1]. We also incorporated image-and-text-aligned pretrained models such as CLIP [29], MetaCLIP [38], DFN-CLIP [12], OpenCLIP [4], DataComp-CLIP [14], ConvNeXT-CLIP [33], EVA02 [13], and SigLIP [39]. Additional image models included SAM-ViT [17] and Inception V3 [36]. For text models, we employed autoencoding models such as RoBERTa [24], BERT [7], ALBERT [19], ModernBERT [37], XLM-RoBERTa [6], as well as sequence-to-sequence models including FLAN-T5 [5] and T5 [30] were used. We also used the text encoders from the aforementioned image-and-text-aligned pretrained models. The complete lists of image and text backbone models used in our experiments are presented in Tables C.1 and C.2, respectively.

B. Results on Text-to-Image

We provide in-depth results on three text-to-image benchmarks: HPDv2 (Tab. B.1), Parti-Prompts (Tab. B.2), and a random selection of COCO prompts (Tab. B.3).

Results on HPDv2. Table B.1 summarizes the rankings and scores for images generated by various models—including one real image—across multiple evaluation metrics. Notably, cFreD achieves the highest alignment with human preferences, reaching a correlation of 0.97. Among statistical metrics, cFreD attains the highest correlation and is comparable to HPSv2 (0.94), a model explicitly trained on human preferences. Given that HPSv2 was trained on the HPSv2 training set, which includes four models from the test set, and employed the same annotators, it inherently encodes specific human preference biases of the same setting. In contrast, cFreD achieves comparable or superior correlation with human evaluation without any human preference training. These results demonstrate that cFreD provides more reliable rankings across diverse models compared to standard automatic metrics and metrics trained explicitly on human preference data.

Table B.1 also reports the rank accuracy scores on the HPDv2 test set. Among all evaluated metrics, cFreD achieves the highest rank accuracy (91.1%), highlighting its strong correspondence with human judgments. HPSv2 follows as the second-best metric with an accuracy of 88.9%, while both FID and $\text{FD}_{\text{DINOv2}}$ obtain competitive scores of 86.7%. Overall, although models trained with human preference data tend to align well with human judgments, cFreD emerges as the most robust and reliable metric.

Result on PartiPrompts Arena. Table B.2 presents the rankings and scores of text-to-image models evaluated on the Parti-Prompt Arena using both statistical metrics and human preference-trained models. Among the statistical metrics, cFreD achieves the highest correlation with human evaluations (0.73), with FID and $\text{FD}_{\text{DINOv2}}$ both reaching a correlation of 0.70. In contrast, the CLIP score shows a very low correlation (0.12) with human judgments. In the human preference trained category, HPSv2 has the strongest alignment, achieving the highest correlation (0.83), followed by ImageReward (0.81) and MPS (0.65). These results highlight that while cFreD is a robust automatic metric, HPSv2 stands out as the most effective in capturing human evaluation trends in the PartiPrompts Arena.

Results on COCO. Table B.3 presents an evaluation on the COCO dataset using nine modern text-to-image models, with human preference rankings sourced from the Text-to-Image Leaderboard and expressed as ELO scores. Among statistical metrics (FID, $\text{FD}_{\text{DINOv2}}$, CLIP, CMMD, and our proposed cFreD), only cFreD exhibits a strong correlation with human preferences, achieving a correlation of 0.33 and a non-trivial rank accuracy of 66.67%. This result places cFreD as the third most aligned metric overall, surpassed only by the human preference–trained metrics ImageReward, HPSv2, and MPS. Notably, all other statistical metrics show considerably weaker alignment with ELO rankings and, as a result, inverted the rankings, resulting in a Rank Acc. below 0.5. These findings highlight that cFreD is sensitive to both visual fidelity and prompt consistency, reinforcing its value as a practical, training-free alternative for benchmarking text-to-image generation.

C. Analysis on Selecting Image and Text Backbone Models

We provide an in-depth analysis of how different image and text models affect cFreD. In our experiments, we tested all possible combinations of 43 different image and text models, focusing exclusively on ViT-based architectures and excluding SigLIP models with high-resolution options.

C.1. Spearman Correlation

We present Spearman correlation heatmaps that compare various text and image models across three distinct datasets: Parti-Prompts, HPDv2, and a random selection of COCO

Models	Statistical Metric												Human Preference Trained Metric							
	Human↑		FID↓		FD_{DINOv2} ↓		CLIP↑		CMMD↓		cFreD↓		Aesthetic↑		ImReward↑		HPS v2↑		MPS↑	
	R#	Rate	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score
GLIDE [25]	1	80.87%	1	7.90	1	6.88	9	14.34	1	2.42	1	3.79	1	5.55	3	0.37	2	25.52	1	12.72
COCO [22]	2	80.66%	7	13.11	7	13.05	10	13.11	5	15.07	4	4.55	5	5.03	1	0.55	1	25.64	5	10.92
FuseDream [23]	3	76.29%	2	8.39	2	7.59	5	15.07	4	5.41	2	4.16	4	5.34	2	0.47	3	24.40	2	12.44
DALLE 2 [31]	4	75.87%	3	9.16	3	7.95	8	14.39	3	4.06	3	4.42	3	5.40	6	0.07	5	23.81	4	11.75
VQGAN+CLIP [10]	5	68.78%	4	10.11	4	8.70	7	14.41	2	3.70	5	4.90	2	5.40	5	0.08	4	23.93	3	11.82
CogView2 [8]	6	39.00%	6	12.65	6	12.87	3	15.45	8	45.64	7	6.93	7	4.82	7	0.02	7	19.45	7	8.85
SDv1.4 [32]	7	38.36%	5	12.51	5	11.93	4	15.42	6	28.52	8	7.18	8	4.56	8	-0.67	8	19.44	8	8.00
VQ-Diffusion [15]	8	32.04%	8	13.85	8	13.12	6	14.71	7	33.40	6	6.59	6	4.88	4	0.17	6	21.91	6	10.15
SDv2.0 [32]	9	22.00%	9	14.74	9	14.23	2	15.62	10	55.88	9	8.16	9	4.54	9	-0.72	9	18.45	9	7.24
LAFITE [40]	10	9.07%	10	15.12	10	14.63	1	16.01	9	53.22	10	9.06	10	4.23	10	-1.45	10	15.03	10	5.08
ρ^2	-		0.70		0.65		0.63		0.88		**0.97**		0.83		0.71		<u>0.90</u>		0.86	
Rank Acc.	-		86.7		86.7		15.6		80.0		**91.1**		82.2		84.4		<u>88.9</u>		86.7	

Table B.1. Text-to-image model ranking and scores by statistical models (FID, FD_{DINOv2}, CLIP score, CMMD, and cFreD) and models that were trained with human preference (Aesthetic Score, ImageReward, HPSv2, and MPS) on HPDv2 test set. Best results in **bold**, second best <u>underlined</u>

Models	Statistical Metric												Human Preference Trained Metric							
	Human↑		FID↓		FD_{DINOv2} ↓		CLIP↑		CMMD↓		cFreD↓		Aesthetic↑		ImReward↑		HPS v2↑		MPS↑	
	R#	Rate	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score
SDXL [28]	1	69.84%	1	31.24	1	23.50	2	32.79	1	2.55	1	2.98	3	5.64	1	0.95	1	28.63	4	10.33
Kand2 [34]	2	46.10%	2	32.08	2	24.35	3	32.62	2	2.77	2	3.21	2	5.65	2	0.90	2	28.13	2	11.29
Wuerst [27]	3	42.68%	3	38.43	3	30.93	4	31.72	3	3.83	3	4.18	1	5.71	3	0.79	3	27.79	1	11.30
Karlo [20]	4	29.21%	4	48.40	4	41.57	1	33.01	4	19.95	4	5.45	4	4.93	4	0.70	4	26.56	3	11.11
ρ^2	-		0.70		0.70		0.12		0.54		0.73		0.43		<u>0.81</u>		**0.83**		0.65	

Table B.2. Text-to-image model ranking and scores by statistical metrics (FID, FD_{DINOv2}, CLIP score, CMMD, and cFreD) and models that were trained with human preference (Aesthetic Score, ImageReward, and MPS) on Parti-Prompt. Best results in **bold**, second best <u>underlined</u>

prompts.

Fig. C.1 showcases the heatmap on Parti-Prompts. Notably, certain image models such as ViT-B/16 trained on ImageNet-1K and ViT-H/14 trained on ImageNet-21K show consistently high performance with different text models. In contrast, SAM-ViT-H/16 presents more variability with different text models. While most ViT, DINO, and CLIP-based models demonstrate strong correlations across different text models, MAE models show slightly lower correlations.

Fig. C.2 presents the heatmap on HPDv2. Most models show strong correlations, whereas MAE models and SAM have lower correlations. Regardless of their performance levels, all image models show consistent correlation patterns across different text models.

Fig. C.3 provides the heatmap on randomly selected COCO prompts. The correlation values are significantly lower overall, ranging from approximately 0.00 to 0.30, which is much lower than both the Parti-Prompts and HPDv2 datasets. Most models show lower correlations, while only DINOv2 models consistently demonstrate stronger correlations. Compared to other datasets, this heatmap exhibits more variability, such as OpenCLIP showing a mixture of relatively higher and lower correlations depending on the text models.

These results highlight the importance of selecting compatible text and image models for improved cross-modal understanding, as not all combinations yield equally robust alignment. Additionally, it suggests that selecting a suitable image encoder plays a more pivotal role than choosing a text encoder, indicating that image encoder choice exerts a greater influence on overall performance.

Models	Statistical Metric												Human Preference Trained Metric							
	Humans↑		FID↓		FD_{DINOv2} ↓		CLIP↑		CMMD↓		cFreD↓		Aesthetic↑		ImReward↑		HPS v2↑		MPS↑	
	R#	ELO	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score
FLUX.1[dev] [18]	1	1083	5	10.45	7	7.21	9	30.72	8	6.08	4	9.93	2	5.75	3	1.10	2	30.74	2	12.90
SDv3.5-L Turbo [11]	2	1073	9	11.68	9	8.12	6	31.11	7	48.52	7	10.44	6	5.50	6	0.64	7	26.52	6	10.88
SDv3.5-L [11]	3	1069	2	10.27	4	6.77	1	31.74	4	40.27	2	9.49	4	5.55	2	1.10	3	30.07	3	12.30
Playgroundv2.5 [21]	4	997	7	10.90	8	7.50	7	31.03	9	66.10	5	10.08	1	6.16	1	1.15	1	31.56	1	13.15
SDv3-M [11]	5	944	6	10.46	5	7.09	3	31.68	2	34.34	1	9.49	7	5.45	4	1.08	4	29.80	9	2.35
SDXL [28]	6	890	3	10.31	2	6.66	2	31.70	6	45.07	3	9.73	3	5.61	5	0.76	5	28.34	4	12.08
SDv2.1 [32]	7	752	1	10.14	1	6.55	4	31.42	3	35.17	6	10.24	5	5.52	8	0.41	6	26.58	5	10.90
Janus Pro [3]	8	740	8	10.97	6	7.17	8	30.99	5	43.51	9	10.76	9	5.33	7	0.57	8	26.22	7	10.70
SDv1.5 [32]	9	664	4	10.41	3	6.73	5	31.21	1	29.89	8	10.58	8	5.34	9	0.19	9	26.15	8	10.50
ρ^2	-		0.08		0.29		0.00		0.22		0.33		0.27		**0.69**		<u>0.48</u>		<u>0.48</u>	
Rank Acc.	-		41.67		36.11		47.22		30.56		66.67		<u>72.22</u>		**80.56**		**80.56**		**80.56**	

Table B.3. Text-to-image model ranking by automatic models(FID, FD_{DINOv2}, CLIP score, CMMD, and cFreD) and models that were trained with human preference (Aesthetic Score, ImageReward, HPSv2, and MPS) on randomly sampled COCO prompts. Rank Acc. below 0.5 indicates there are more discordant pairs than concordant ones. Best results in **bold**, second best <u>underlined</u>

C.2. Effect of Visual Encoders on HPDv2

In this section, we provide an analysis of how different visual encoder characteristics impact cFreD's correlation with human preferences on HPDv2. For each factor examined, we report the average correlation across all possible text encoder combinations.

Effect of the size of the pre-training dataset on cFreD**.** Fig. C.4a shows the correlation between cFreD and human preferences as a function of the size of the pre-training dataset for Vision Transformer (ViT). In all ranges, it show high alignment with human judgments, with a correlation higher than 0.95 correlation across all data sizes. This indicates that factors beyond raw data quantity, such as diversity and quality, significantly influence performance.

Effect of image size on cFreD**.** Fig. C.4b illustrates the correlation between cFreD and human preferences across varying input image resolutions. We observe a nonmonotonic relationship: increasing resolutions do not consistently yield higher correlations. In particular, an image size from 224x224 to 896x896 all achieves a high correlation abovethan 0.96. However, beyond 518x518, performance declines, reaching 0.964 at 896x896 and showing the lowest correlation of 0.774 at 1024x1024.

Effect of ViT model size on cFreD**.** The correlation between cFreD and human preferences across Vision Transformer (ViT) sizes are presented in Fig. C.4c. It shows that all models achieve consistently high correlations—ranging from 0.945 to 0.986—indicating strong agreement with the target metric across scales. Interestingly, the *SO* model attains the highest correlation at 0.986, while *Gigantic* has the lowest, though still robust correlation of 0.945. The remaining models also cluster around correlations between 0.96 and 0.98, suggesting that simply increasing model size does not guarantee a strictly monotonic improvement in correlation.

Effect of ViT Feature dimensionality on cFreD**.** Fig. C.4d shows the correlation between cFreD and human preferences across ViT feature dimensions from 256 to 1664. The lowest number of the feature (256) shows the lowest correlation. However, we observe a clear plateau effect in performance once the feature count reaches 384, with correlation values stabilizing around 0.98-0.99 across a wide range of dimensionalities (512-1408). Interestingly, at extremely high dimensionalities (above 1536), we note a slight performance decline, with correlation dropping to 0.95 at 1664 features. This suggests an optimal range for feature dimensionality exists, beyond which additional computational complexity yields diminishing or even negative returns.

Effect of Zero-Shot ImageNet Accuracy on cFreD**.** Fig. C.4e depicts a boxplot of the correlation between cFreD and human preferences as a function of zero-shot ImageNet accuracy, evaluated exclusively on image-text pretrained models [29]. Higher zero-shot accuracies generally correspond to stronger correlations with human judgments, though variance exists within each accuracy bin. Interestingly, we find that correlations peak at a model with 66.58% zero-shot accuracy and decrease as model accuracy gets higher.

C.3. Effect of Visual Encoders on COCO prompts

In this section, we provide an analysis of how different visual encoder characteristics impact cFreD's correlation with human preferences on randomly selected COCO prompts. For each factor examined, we report the average correlation across all possible text encoder combinations.

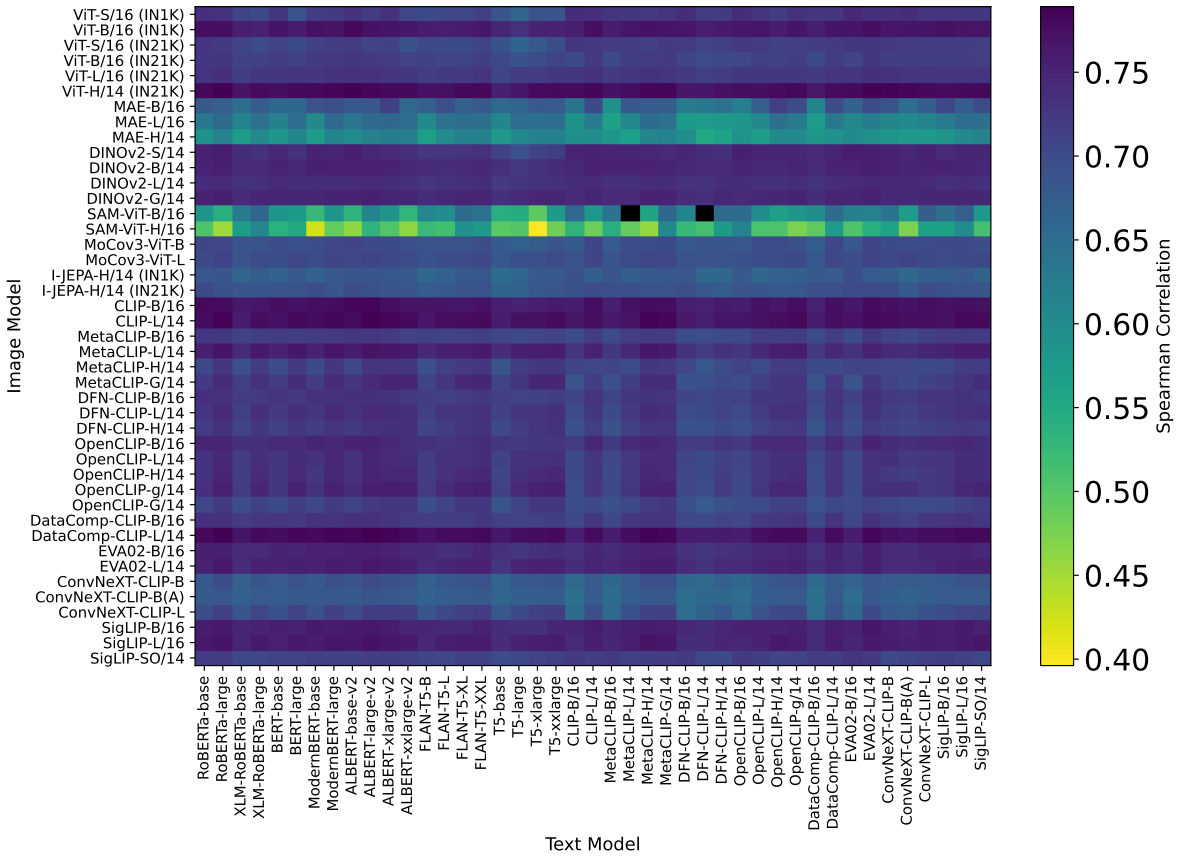


Figure C.1. Spearman Correlation Heatmap on Parti-Prompts.

Effect of the size of the pre-training dataset on cFreD**.** Fig. C.5a illustrates the correlation between cFreD and human preferences as a function of the pre-training dataset size for Vision Transformer (ViT). We observe a nonmonotonic relationship: increasing the pre-training dataset size does not consistently yield higher correlations. Notably, a model trained with fewer than 100 million samples achieves a high correlation of 0.33, while models trained with larger datasets show lower correlations ranging from 0.06 to 0.21. Interestingly, models trained with fewer than 5 billion samples demonstrate the lowest human correlation (0.06). These findings indicate that simply scaling up data does not guarantee improved performance for every metric or task, suggesting that balancing the quantity and quality of training data is crucial for optimal results.

Effect of image size on cFreD**.** Fig. C.4b illustrates the correlation between cFreD and human preferences across varying input image resolutions. We observe an inconsistent relationship: starting around 0.05 at an image size of 224x224, then rising to about 0.10 at 256x256 before dropping again to around 0.03 at 299x229. After a modest increase to 0.08 at 384x384, the correlation dips to 0.01 at 448x448 and then peaks sharply at 0.18 for an image size of 518x518. Beyond that, it decreases to nearly 0 at 896x896 and recovers slightly to 0.03 at 1024. These erratic trends suggest that there is no straightforward, monotonic relationship between image size and correlation for this particular task or metric, and the highest correlation appears in the midrange rather than at the smallest or largest resolutions.

Effect of ViT model size on cFreD**.** The correlation between cFreD and human preferences across Vision Transformer (ViT) sizes are presented in Fig. C.4c. Larger models tend to improve alignment with human judgments, with *Giant* model achieving the highest correlation of 0.320. However, when the model size gets bigger to *Gigantic* model, the correlation degrades down to 0.089. These results indicate that correlation does not simply increase in tandem with model size; rather, there seems to be an optimal range, as exemplified by the *Giant* model, for achieving the strongest alignment with the evaluation metric.

Effect of ViT Feature dimensionality on cFreD**.** Fig. C.4d shows the correlation between cFreD and human preferences across ViT feature dimensions from 256 to 1664. The two lowest numbers of the features (256 and 384) show the lowest correlation. However, we observe a clear plateau effect in performance once the feature count reaches 512,

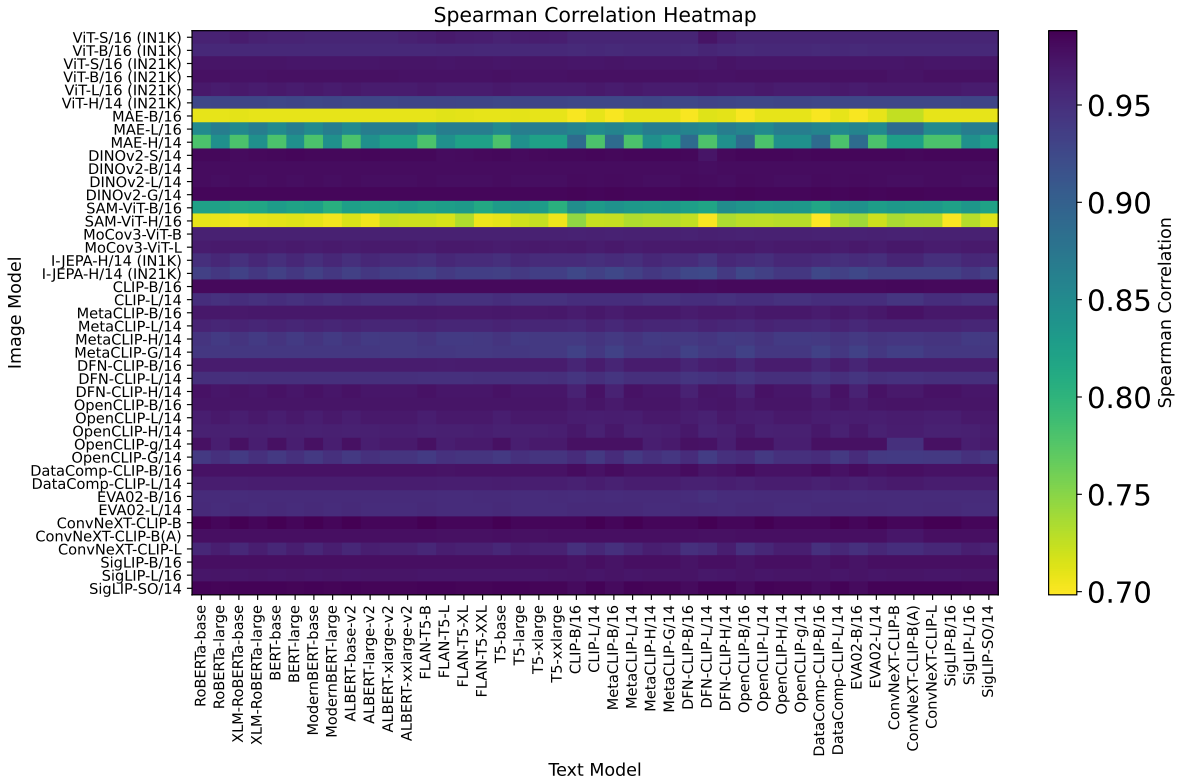


Figure C.2. Spearman Correlation Heatmap on HPDv2.

with correlation values stabilizing around 0.30-0.32 across a wide range of dimensionalities (512-1152). Interestingly, at extremely high dimensionalities (above 1536), we note a slight performance decline, with correlation dropping to 0 at the 1408 feature. Although the correlation increases back to 0.32 with 1536 features, it decreases back to 0.14 at 1664 features. This suggests an optimal range for feature dimensionality exists, beyond which additional computational complexity yields diminishing or even negative returns.

Effect of Zero-Shot ImageNet Accuracy on cFreD**.** Fig. C.4e depicts a boxplot of the correlation between cFreD and human preferences as a function of zero-shot ImageNet accuracy, evaluated exclusively on image-text pretrained models [29]. Interestingly, we find that the lowest zero-shot accuracy has the highest correlation to human preference. However, after 68.58 of zero-shot accuracies, higher zero-shot accuracies generally correspond to stronger correlations with human judgments, though variance exists within each accuracy bin.

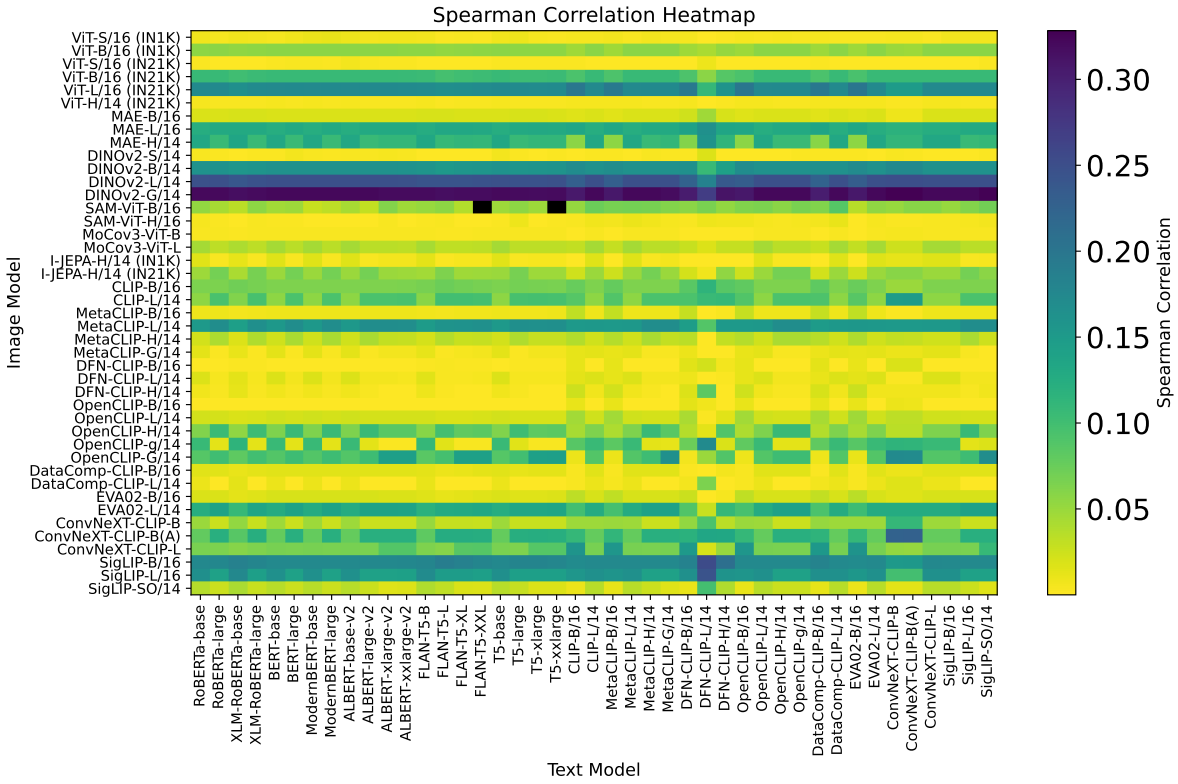
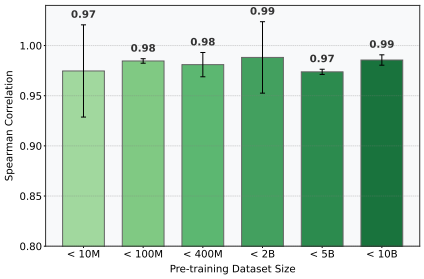
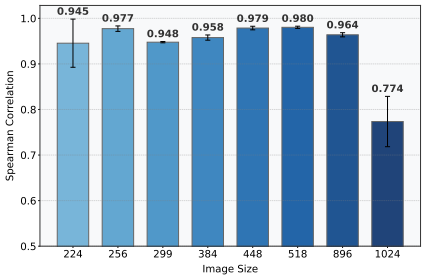


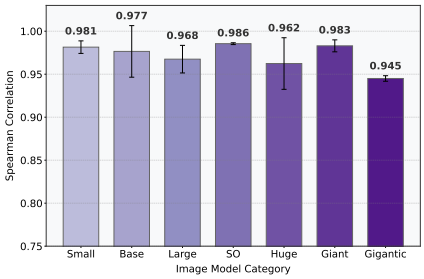
Figure C.3. Spearman Correlation Heatmap on random COCO prompts.



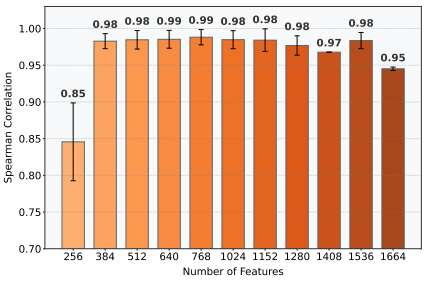
(a) Effect of training data.



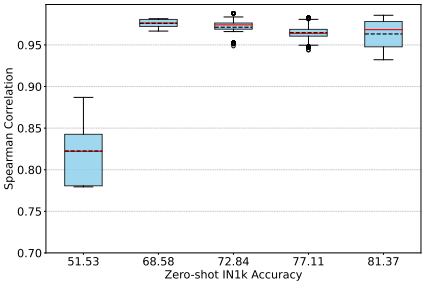
(b) Effect of image size.



(c) Effect of model size.

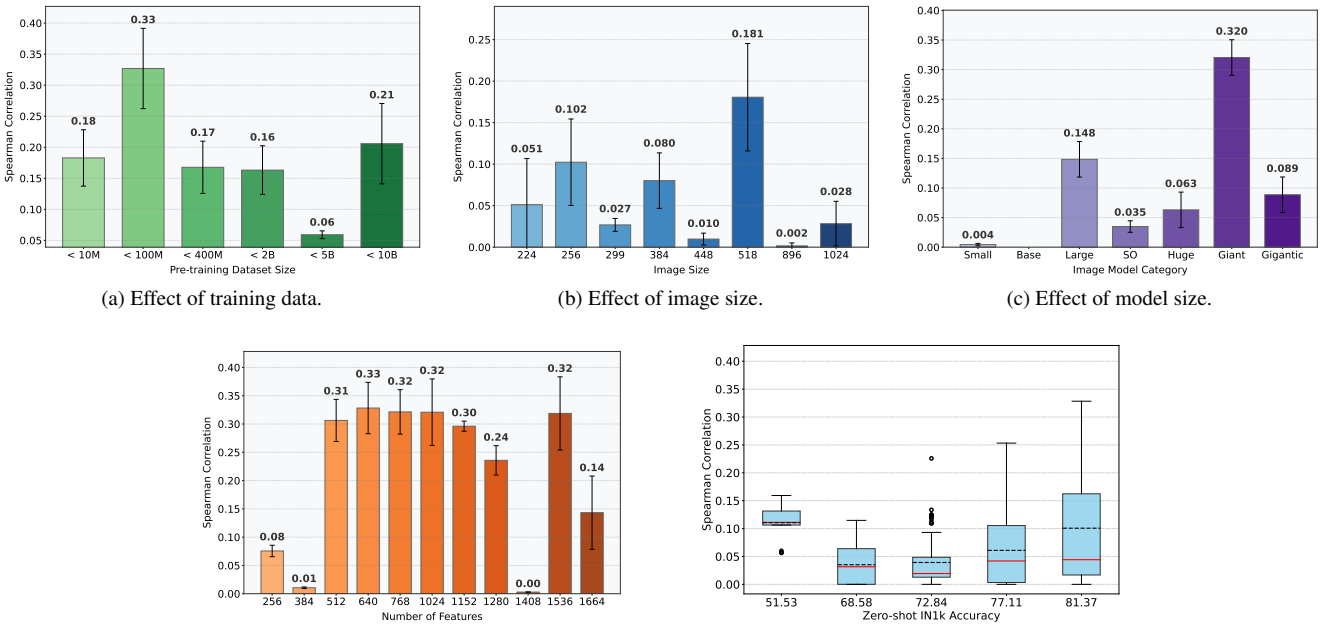


(d) Effect of number of features.



(e) Effect of IN1K Zero-shot accuracy.

Figure C.4. **Ablation study on the HPDv2** dataset comparing the correlation to human preferences under varying factors: (a) the ViT training dataset, (b) input image size, (c) model capacity, (d) the number of features in the last ViT layer, and (e) zero-shot accuracy on ImageNet-1K.



(a) Effect of training data. (b) Effect of image size. (c) Effect of model size.

(d) Effect of number of features. (e) Effect of IN1K Zero-shot accuracy.

Figure C.5. **Ablation study on randomly sampled COCO prompts** comparing the correlation to human preferences under varying factors: (a) the ViT training dataset, (b) input image size, (c) model capacity, (d) the number of features in the last ViT layer, and (e) zero-shot accuracy on ImageNet-1K.

Model Name	Model Path
ViT-S/16 (IN1K)	timm/vit_small_patch16_224.augreg_in1k
ViT-B/16 (IN1K)	timm/vit_base_patch16_224.augreg_in1k
ViT-S/16 (IN21K)	timm/vit_small_patch16_224.augreg_in21k
ViT-B/16 (IN21K)	timm/vit_base_patch16_224.augreg_in21k
ViT-L/16 (IN21K)	timm/vit_large_patch16_224.augreg_in21k
ViT-H/14 (IN21K)	timm/vit_huge_patch14_224.orig_in21k
MAE-B/16	timm/vit_base_patch16_224.mae
MAE-L/16	timm/vit_large_patch16_224.mae
MAE-H/14	timm/vit_huge_patch14_224.mae
DINOv2-S/14	timm/vit_small_patch14_dinov2.lvd142m
DINOv2-B/14	timm/vit_base_patch14_reg4_dinov2.lvd142m
DINOv2-L/14	timm/vit_large_patch14_dinov2.lvd142m
DINOv2-G/14	timm/vit_giant_patch14_dinov2.lvd142m
SAM-ViT-B/16	timm/samvit_base_patch16
SAM-ViT-H/16	timm/samvit_huge_patch16
MoCov3-ViT-B	nyu-visionx/moco-v3-vit-b
MoCov3-ViT-L	nyu-visionx/moco-v3-vit-l
I-JEPA-H/14 (IN1K)	jmtzt/ijepa_vith14_1k
I-JEPA-H/14 (IN21K)	facebook/ijepa_vith14_22k
CLIP-B/16	timm/vit_base_patch16_clip_224.openai
CLIP-L/14	timm/vit_large_patch14_clip_224.openai
MetaCLIP-B/16	timm/vit_base_patch16_clip_224.metaclip_400m
MetaCLIP-L/14	timm/vit_large_patch14_clip_224.metaclip_400m
MetaCLIP-H/14	timm/vit_huge_patch14_clip_224.metaclip_altogether
MetaCLIP-G/14	timm/vit_gigantic_patch14_clip_224.metaclip_2pt5b
DFN-CLIP-B/16	timm/vit_base_patch16_clip_224.dfn2b
DFN-CLIP-L/14	timm/vit_large_patch14_clip_224.dfn2b
DFN-CLIP-H/14	timm/vit_huge_patch14_clip_224.dfn5b
OpenCLIP-B/16	timm/vit_base_patch16_clip_224.laion2b
OpenCLIP-L/14	timm/vit_large_patch14_clip_224.laion2b
OpenCLIP-H/14	timm/vit_huge_patch14_clip_224.laion2b
OpenCLIP-g/14	timm/vit_giant_patch14_clip_224.laion2b
OpenCLIP-G/14	timm/vit_gigantic_patch14_clip_224.laion2b
DataComp-CLIP-B/16	timm/vit_base_patch16_clip_224.datacompxl
DataComp-CLIP-L/14	timm/vit_large_patch14_clip_224.datacompxl
EVA02-B/16	timm/eva02_base_patch16_clip_224
EVA02-L/14	timm/eva02_large_patch14_clip_224
ConvNeXT-CLIP-B	timm/convnext_base.clip_laion2b
ConvNeXT-CLIP-B(A)	timm/convnext_base.clip_laiona
ConvNeXT-CLIP-L	timm/convnext_large_mlp.clip_laion2b_augreg
SigLIP-B/16	timm/vit_base_patch16_siglip_224.webli
SigLIP-L/16	timm/vit_large_patch16_siglip_256.webli
SigLIP-SO/14	timm/vit_so400m_patch14_siglip_gap_224.webli
InceptionV3	inception/inceptionv3
SigLIP-SO/14 (448)	timm/vit_so400m_patch14_siglip_gap_448.pali2_10b_pt
SigLIP-SO/14 (896)	timm/vit_so400m_patch14_siglip_gap_896.pali2_10b_pt

Table C.1. List of image backbone models used in our analysis.

Model Name	Model Path
RoBERTa-base	FacebookAI/roberta-base
RoBERTa-large	FacebookAI/roberta-large
XLM-RoBERTa-base	FacebookAI/xlm-roberta-base
XLM-RoBERTa-large	FacebookAI/xlm-roberta-large
BERT-base	google-bert/bert-base-uncased
BERT-large	google-bert/bert-large-uncased
ModernBERT-base	answerdotai/ModernBERT-base
ModernBERT-large	answerdotai/ModernBERT-large
ALBERT-base-v2	albert/albert-base-v2
ALBERT-large-v2	albert/albert-large-v2
ALBERT-xlarge-v2	albert/albert-xlarge-v2
ALBERT-xxlarge-v2	albert/albert-xxlarge-v2
FLAN-T5-B	google/flan-t5-base
FLAN-T5-L	google/flan-t5-large
FLAN-T5-XL	google/flan-t5-xl
FLAN-T5-XXL	google/flan-t5-xxl
T5-base	google/t5-v1_1-base
T5-large	google/t5-v1_1-large
T5-xlarge	google/t5-v1_1-xl
T5-xxlarge	google/t5-v1_1-xxl
CLIP-B/16	ViT-B-16-quickgelu.openai
CLIP-L/14	ViT-L-14-quickgelu.openai
MetaCLIP-B/16	ViT-B-16-quickgelu.metaclip_400m
MetaCLIP-L/14	ViT-L-14-quickgelu.metaclip_400m
MetaCLIP-H/14	ViT-H-14.metaclip_altogether
MetaCLIP-G/14	ViT-bigG-14-CLIPA.datacomp1b
DFN-CLIP-B/16	ViT-B-16-quickgelu.dfn2b
DFN-CLIP-L/14	ViT-L-14-quickgelu.dfn2b
DFN-CLIP-H/14	ViT-H-14-quickgelu.dfn5b
OpenCLIP-B/16	ViT-B-16.laion2b_s34b_b88k
OpenCLIP-L/14	ViT-L-14.laion2b_s32b_b82k
OpenCLIP-H/14	ViT-H-14.laion2b_s32b_b79k
OpenCLIP-g/14	ViT-g-14.laion2b_s34b_b88k
DataComp-CLIP-B/16	ViT-B-16.datacomp_xl_s13b_b90k
DataComp-CLIP-L/14	ViT-L-14.datacomp_xl_s13b_b90k
EVA02-B/16	EVA02-B-16.merged2b_s8b_b131k
EVA02-L/14	EVA02-L-14.merged2b_s4b_b131k
ConvNeXT-CLIP-B	convnext_base_w.laion2b_s13b_b82k_augreg
ConvNeXT-CLIP-B(A)	convnext_base_w.laion_aesthetic_s13b_b82k
ConvNeXT-CLIP-L	convnext_large_d.laion2b_s26b_b102k_augreg
SigLIP-B/16	ViT-B-16-SigLIP.webli
SigLIP-L/16	ViT-L-16-SigLIP-256.webli
SigLIP-SO/14	ViT-SO400M-14-SigLIP.webli

Table C.2. List of text backbone models used in our analysis.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 1

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 1

[3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 3

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 1

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 1

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1

[8] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[12] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 1

[13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023. 1

[14] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023. 1

[15] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 2

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[18] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3

[19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 1

[20] Donghoon Lee, Jiseob Kim, Jisu Choi, Jongmin Kim, Minwoo Byeon, Woonhyuk Baek, and Saehoon Kim. Karlo-v1.0.alpha on coyo-100m and cc15m. https://github.com/kakaobrain/karlo, 2022. 2

[21] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. 3

[22] 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

[23] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021. 2

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized

bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[27] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 3, 5

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 1

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

[34] Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. kandinsky 2.2, 2023. 2

[35] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 1

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[37] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024. 1

[38] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 1

[39] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[40] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17907–17917, 2022. 2