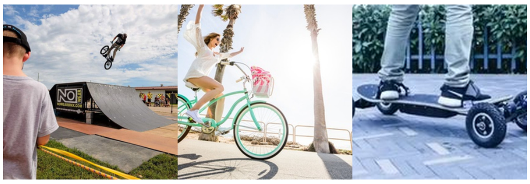
The Association Between Training Data and Success Ratios in Text-to-Image Generation

Anonymous ACL submission

Abstract

Text-to-image (T2I) models are often touted for their supposed ability to create compositional images with many components. However, these models can fail to generate all entities when presented with prompts containing just two or three entities. In this work, we seek an explanation of such failures with respect to the training data. We introduce the *training appearance ratio*, which compares the number of training images depicting specific entities vs. the number of training captions mentioning those same entities, and examine how well this measure correlates with generation success rates. We find positive and significant correlations between these ratios and successful image generations. Furthermore, our proposed measure yields stronger correlations with model success rates than existing training data frequency measures. These associations suggest that our measure (*training appearance ratio*) better captures the relationship between training data and generation success.

(a) Generated images for the prompt "a **bicycle** and a **skateboard**". The model (SD1.5) mostly generates *one of the two* objects (primarily bicycles).



(b) Training images where *either* **skateboard** or **bicycle** are shown, *but not both*. Many of these images depict parks and outdoors spaces that are suitable for both skateboarding and bicycling, but only include one.

Figure 1: Examples of generated/training images where prompts/captions mention "skateboard" and "bicycle", but corresponding images do not include both.

1 Introduction

When asked to generate an image of "a bicycle and a skateboard", Stable Diffusion, a popular text-to-image (T2I) model (Rombach et al., 2022), succeeds only 8% of the time. Despite "bicycle" and "skateboard" being common objects that are generated separately nearly 100% of the time, the model fails to generate both jointly. The inability of models to handle such simple cases showcase their weak compositional capabilities.

In this work, we aim to explain models' failures with respect to their training data properties. Drawing from previous works that have shown that pretraining data frequencies correlate with model performance (Razeghi et al., 2022; Kandpal et al., 2023; Udandarao et al., 2024), we first seek to replicate such findings for our setup of generating multiple common entities. However, our results indicate that simple caption frequencies correlate poorly with models' generation success rates. Upon digging into the training data, we observe that captions mentioning entities may pair with images that only showcase a subset of those entities, or none at all, as shown in Figure 1b. For instance, there are more than 9,000 captions in LAION2B-en (Schuhmann et al., 2022) that mention both "bicycle" and "skateboard", but only 9% of corresponding images actually contain both objects. These findings indicate that captions alone provide an inaccurate measure of how often entities are actually depicted in training images.

Based on these findings, we adjust our frequencies to only consider training examples for which both the captions and images contain all specified entities (Udandarao et al., 2024). While these adjusted frequencies correlate better with models' generation success rates, they do not account for how T2I models are trained and used in practice (i.e., images are conditioned on texts). Therefore, we consider the ratio between entities appearance in training images vs. captions, which explicitly incorporates this conditioning, and formalize this measure to be the *training appearance ratio*. We find that this ratio exhibits stronger correlations with models' generation capabilities across various combinations of models, prompts, and entities ($\rho = 0.43$ vs. 0.27 for 2 entities, and $\rho = 0.31$ vs. 0.19 for 3 entities, averaged). These stronger correlations show that our measure better associates success in generating images with the training data.

In summary, our work demonstrates that models are poor at basic compositional generations, and proposes a new training data measure that correlates better with models' success rates than existing approaches. Our findings suggest that simple training appearance ratios can help better understand model behavior, and lay the foundation for future work that investigates concrete explanations for model failures and successes.

2 Explaining Successes Through Training Data Statistics

T2I models often fail to generate images following simple prompts with multiple common entities. Our main goal in this study is to investigate whether models' ability to faithfully generate images based on prompts can be attributed to statistics from their training data. To address this objective, we need to first define how we measure and compare training data statistics and image generation success. Consider a training dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ consisting of N (image, caption) pairs. We also assume a prompt p that instructs the model to generate some entities $\boldsymbol{e} = \{e_1, e_2, ..., e_k\}$, where $\forall i, e_i \in p$. To identify relevant examples from $\mathcal{D}$ we select training captions that mention the entities $\boldsymbol{e}$ specified in p. For example, for the prompt "a bicycle and a skateboard", we query from $\mathcal{D}$ and choose image-caption pairs whose captions include the entities "bicycle" and "skateboard".

Note that while entities $\boldsymbol{e}$ may appear in a caption y_i, the image x_i corresponding to that caption may not contain all entities (sometimes even none), as depicted in Figure 1, and as was observed in Udandarao et al. (2024).[1] Since raw counts provide a biased estimation of entity occurrences in images, we propose measuring the proportion of captions whose images also contain all specified entities.

[1] Table 5 (Appendix) shows example image-caption pairs.

We define this quantity to be the training appearance ratio ($tar_{e,ic}$):

$$tar_{e,ic} = \frac{|\mathcal{D}_{e,i}|}{|\mathcal{D}_{e,c}|}$$

where $\mathcal{D}_{e,c}$ is the subset of $\mathcal{D}$ whose captions contain entities $\boldsymbol{e}$, and $\mathcal{D}_{e,i}$ is the subset of $\mathcal{D}$ whose captions and images contain entities $\boldsymbol{e}$. A higher value of $tar_{e,ic}$ indicates that image-caption pairs that mention a set of entities in captions also tend to include those entities in images.

After computing $tar_{e,ic}$, we generate images for prompt p using a T2I model to obtain generated images $\mathcal{G}_{e,p}$. We calculate the proportion of images that depict all entities, which we call the generation appearance ratio ($gar_{e,ip}$).

$$gar_{e,ip} = \frac{|\mathcal{G}_{e,i}|}{|\mathcal{G}_{e,p}|}$$

Similar to above, $\mathcal{G}_{e,i}$ is the subset of generated images whose prompts and images contain entities $\boldsymbol{e}$. We then examine whether the generation appearance ratio of generated entities that are explicitly specified in prompts ($gar_{e,ip}$) correlates with corresponding ratios from the training data ($tar_{e,ic}$). While previous works highlight correlations between model behavior and frequencies in the data (Razeghi et al., 2022; Kandpal et al., 2023; Udandarao et al., 2024), we hypothesize that training appearance ratios exhibit stronger associations with model generation capabilities, since $tar_{e,ic}$ directly captures discrepancies in how often entities occur in training images vs. texts (similar to how $gar_{e,ip}$ captures discrepancies in how often entities occur in generated images vs. prompts). In other words, we argue that $tar_{e,ic}$ more closely matches what we measure at generation, resulting in stronger correlations as we show in Section 4.

3 Experimental Setup

Entities We select entities from the MS COCO dataset (Lin et al., 2014) classes in addition to manually added entities (e.g., fruits, vegetables) as shown in Table 3 (Appendix), resulting in 84 entities. We intentionally focus on frequent entities that models succeed in generating individually.

Automated Image Evaluation To determine whether an image contains specified entities, we utilize an automated approach. We use visual question answering (VQA) and employ PaliGemma

(Google, 2024) as our VQA model. More specifically, we ask the model whether an image contains a given entity, which is done for all entities in the prompt, and consider an image to contain all entities if the model answers "yes" for every entity. Note that PaliGemma achieves 91% on human annotated images, as discussed in Appendix A.5.

Entity Caption Occurrences We use WIMBD (Elazar et al., 2024) to retrieve counts of entities from the training data. Specifically, we extract captions that mention a set of entities ($\mathcal{D}_{e,c}$), and randomly sample up to $1,000$ image-caption pairs. Based on the corresponding images, we calculate the proportion of images that depict the specified entities to measure $tar_{e,ic}$. We multiply the number of captions ($|\mathcal{D}_{e,c}|$) by the ratios computed previously, $tar_{e,ic}$, to estimate the number of training examples that both mention entities in captions and include them in images.

Prompts We prompt the model to generate images with one, two, and three entities using the prompts shown in Table 4 in Appendix A.2. For each prompt, we generate 50 images using different random seeds, resulting in 100 images total for single entity prompts and 200 images total for double and triple entity prompts.

Data & Models We focus on Stable Diffusion (Rombach et al., 2022), a popular set of text-to-image models. Specifically, we use SD1.1 and SD1.5, which are both trained on 2.3 billion image-caption pairs filtered to contain only English captions (LAION2B-en). Additionally we use SD2.1, which is trained on LAION-5B (Schuhmann et al., 2022), a dataset of 5.9 billion multilingual image-captions pairs (including LAION2B-en). Notably, these are the only two T2I training datasets indexed in the WIMBD tool, which is important because working with such massive datasets without proper tooling is incredibly challenging.

4 Results

Generation Appearance Ratios How good are models at compositional generation? To answer this question, we examine generation appearance ratios ($gar_{e,ip}$), which capture the success rate of generating images with all specified entities, for different models and number of entities (Table 1). We find that all models successfully generate single entities $> 96\%$ of the time, validating that models are capable of generating common individual entities.

Model	1 Entity	2 Entities	3 Entities
SD1.1	0.98	0.44	0.18
SD1.5	0.99	0.50	0.21
SD2.1	0.96	0.66	0.32

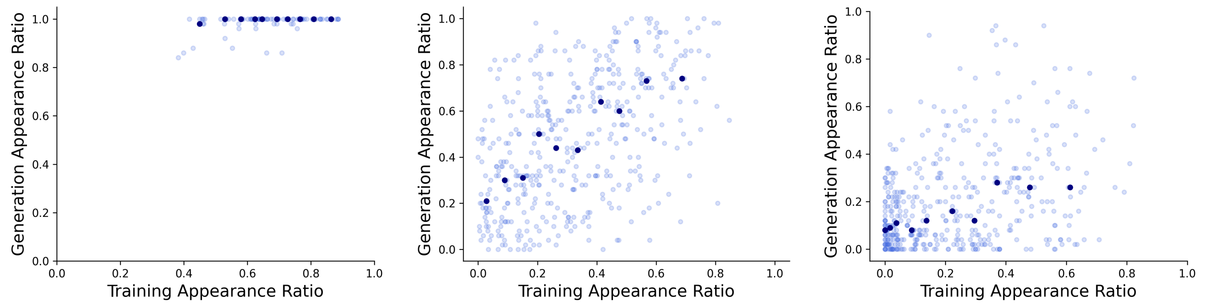
Table 1: Generation appearance ratios ($gar_{e,ip}$) for different models and # of entities, averaged across prompts.

However, models exhibit massive drops when generating two and three entities – for example, both SD1.1 and SD1.5 models generate two entities $<=$ 50% of the time. Although SD2.1 is notably better at generating two entities (at nearly 66%), it still struggles in this simple compositional setting. In summary, we see that models fail increasingly as prompts depict more entities. We do not go beyond 3 entities, since Stable Diffusion generates four entities $< 5\%$ of the time.

Correlations between Model Behavior and Training Data Statistics We wish to explain model success rates in generating various entities with respect to the training data. To do so, we first analyze frequency-based approaches, building on related work that explores the impact of training data in different settings (Razeghi et al., 2022; Udandarao et al., 2024). We then show that our proposed measure ($tar_{e,ic}$) is more strongly correlated with model behavior.

Baselines: Frequency-based Approaches As baselines, we compute Pearson's correlation between $gar_{e,ip}$ and (1) frequencies of entities in captions and (2) estimated frequencies of entities in images (counts multiplied by $tar_{e,ic}$). Following Udandarao et al. (2024), we compute the log_{10} of frequencies to capture log-linear associations, and refer to the resulting correlations as ρ_{cap} and ρ_{im}. Results are presented in the first two sections of Table 2 for various models and number of entities, averaged across prompts.

We find that ρ_{cap} is not statistically significant (significance level < 0.01) across all combinations of models, prompts, and number of entities except for SD1.1 with one entity. For the overwhelming majority of cases, raw captions counts do not correlate with $gar_{e,ip}$. These results are unsurprising, since raw caption counts are poor indicators of how often entities actually occur in training images. We observe negative correlations for ρ_{cap} in the three entity case, which is somewhat surprising, but these values are not statistically significant. In contrast, ρ_{im} exhibits consistently positive correlations for



(a) 1 Entity (b) 2 Entities (c) 3 Entities

Figure 2: Correlations between generation appearance ratios ($gar_{e,ip}$) and training appearance ratios ($tar_{e,ic}$) for 1, 2, and 3 entities, shown for SD1.1 and prompt 1. We bin examples into 10 equally-sized groups or deciles based on $tar_{e,ic}$ and compute median $tar_{e,ic}$ and $gar_{e,ip}$ values for each bin, which correspond to the navy blue points.

two and three entities, and is statistically significant across all prompts and models in the two entity case. When comparing ρ_{im} values for two and three entities, we observe a clear reduction in ρ_{im} across models (0.08 absolute decrease). This reduction may be due to models exhibiting poor generation capabilities as a whole for three entities. Overall, these findings indicate that frequency-based measures may not be effective in capturing the generation success for multiple entities.

Proposed Measure: Training Appearance Ratios We present correlation results between $gar_{e,ip}$ and $tar_{e,ic}$ in the last section of Table 2 (ρ_{ratio}). We find that all models exhibit positive, but not statistically significant correlations for single entities. Since we select frequently occurring entities by design, we can expect models to generate them successfully irrespective of $tar_{e,ic}$.

For prompts with two and three entities, we observe positive and statistically significant correlations across all models, prompts, and number of entities. Both Figures 2b (two entities, and 2c (three entities) show linear associations between generation and training appearance ratios. These associations become much clearer when data points are binned into deciles based on $tar_{e,ic}$, with ρ_{ratio}=0.95 for 2 entities and ρ_{ratio}=0.90 for 3 entities. We observe some variability across prompts with $\sigma \leq 0.07$ for two entities and $\sigma \leq 0.06$ for three entities. Similar to ρ_{im}, we see a decrease in ρ_{ratio} going from two to three entities (0.12 absolute decrease). That being said, ρ_{ratio} consistently exhibits statistical significance and higher values relative to ρ_{im}. Overall, these results suggest ρ_{ratio} is a stronger indicator of successful generations for compositional prompts depicting multiple entities.

Corr	Model	Number of Entities		
		1	2	3
ρ_{cap}	SD1.1	**0.37	0.06	-0.12
	SD1.5	0.12	0.07	-0.06
	SD2.1	0.20	0.02	-0.06
ρ_{im}	SD1.1	**0.40	**0.31	0.20
	SD1.5	0.18	**0.28	0.17
	SD2.1	0.26	**0.23	0.21
ρ_{ratio} (ours)	SD1.1	0.17	**0.47	**0.34
	SD1.5	0.29	**0.42	**0.28
	SD2.1	0.23	**0.40	**0.30

Table 2: Pearson's correlation coefficients between generation appearance ratios and various training data measures: (1) frequency of entities in captions (ρ_{cap}) as a baseline, (2) estimated frequency of entities in images (ρ_{im}) as another baseline, and (3) our proposed measure (ρ_{ratio}), averaged across prompts. We compute the $\log_{10}$ of frequencies for (1) and (2) to capture log-linear associations. ** indicates correlations are statistically significant (significance level < 0.01) for all prompts.

5 Conclusion

This work studies the connection between models' generation success and *training appearance ratios*. Although numerous studies have shown that model performance strongly correlates with the frequency of entities (Razeghi et al., 2022; Kandpal et al., 2023; Udandarao et al., 2024), we show that for image generation, successful generations correlate better with our proposed ratios. Our findings are complemented by Seshadri et al. (2023), who also show that model generations are associated with ratios from the training data in the context of gender-occupation biases. Our results emphasize the need for improving data quality by limiting image-caption mismatches and further necessitates open access to pretraining corpuses to be able to characterize model behaviors and their flaws.

Limitations

We compare properties in the training data with model behavior using correlational analysis and observe clear trends: higher training appearance ratios are associated with higher generation successes. However, we cannot assert that our measure explains or definitively impacts model behavior without employing a causal approach, and leave this important direction to future work.

Our results suggest that different entity combinations with similar training appearance ratios can have variable generation success rates. Although correlations between training appearance ratios and model success rates are consistently positive and significant in the two and three entity settings, they are weakly to moderately positive. These results suggest that simple training appearance ratios offer some insights into models' generation capabilities, but do not provide the full story. Perhaps there are more nuanced training data measures to consider, or other factors beyond the data such as model scale, architecture, and training.

Along these lines, it is worth noting that closed models such as DALL-E 2 (Ramesh et al., 2022), and especially DALL-E 3 (Betker et al., 2023), are much better at handling compositional prompts. While we do not know the exact factors that contribute to this improvement, we speculate that training data quality and curation play a huge role. Perhaps the image-caption pairs used to train such models were filtered or augmented to have much higher training appearance ratios as a whole. However, without access to such datasets, it is unclear whether training appearance ratios are a driving force behind more capable models.

In addition, we focus on the specific setup of generating between 1-3 entities, which is a fundamental aspect of compositional understanding. As we show, models fail considerably even in this simple setting. However, there are other well-known failure modes (Ghosh et al., 2023; Huang et al., 2023; Rassin et al., 2023) in text-to-image generation that should be considered. Furthermore, our study focuses exclusively on English prompts. We encourage researchers to study the association between training data and text-to-image generation for other languages. This study is among the first to investigate text-to-image failure modes with respect to training data. We hope that this study motivates future work to further probe and expand on these findings.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Preprint*, arXiv:2310.11513.

Google. 2024. Big vision: Paligemma project configurations. `https://github.com/google-research/big_vision/tree/main/big_vision/configs/proj/paligemma`. Accessed: 2024-06-14.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20406–20417.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv: 2307.06350*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685. IEEE Computer Society.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. 2023. The bias amplification paradox in text-to-image generation. *ArXiv*, abs/2308.00755.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H.S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2024. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *ArXiv*, abs/2404.04125.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Prompts

The prompts used for generating images are presented in Table 4. For each prompt, we have the following number of instances (i.e., entity combinations after filling in [E1], [E2], [E3]): we have 84 instances for 1 entity, 440 instances for 2 entities, and 440 instances for 3 entities.

A.2 Image Generation

This work uses 3 Stable Diffusion versions: SD1.1 and SD1.5 (trained on LAION2B-en) and SD2.1 (trained on LAION-5B). We use the default generation parameters of 50 inference steps and a guidance scale of 7.5. We specify a batch size of 4. For a given instance of a prompt (i.e., filled in with entities) and model version, we generate 50 images using different random seeds. In total, our generations have taken ~600 hours in total on a single TITAN RTX GPU.

A.3 Entities

The entities used to fill in prompts are presented in Table 3. We include 84 entities in total. The minimum count in in the dataset is for the word "beet" with 123,134 caption mentions for LAION2B-en and 194,530 caption mention for LAION5B. The maximum count is for the word "book" with 21,353,659 caption mentions in LAION2B-en and 28,379,268 for LAION5B.

A.4 VQA

For performing automated image evaluation, a common choice is to use CLIPScore (Hessel et al., 2021). However, CLIP (Radford et al., 2021), its underlying model, struggles with compositional understanding (Hu et al., 2023; Yuksekgonul et al., 2023) and performs poorly for such prompts. As a result, we turn to Visual Question Answering (VQA). We ask a separate question for each entity using the following format: "Is there a/an [entity] in this image, yes or no?", which is then asked for all entities in the prompt. If the model responds "yes" to each of the questions, we consider the image to contain all specified entities. This approach is used for both training and generated images.

A.5 Human Evaluation

We perform human evaluation to assess whether our VQA approach is appropriate and effective for evaluating the presence of entities in images. The

Entities				
airplane	apple	asparagus	backpack	banana
bear	bed	beet	bench	bicycle
bird	boat	book	bottle	bowl
broccoli	bus	cake	car	carrot
cat	chair	clock	coconut	corn
couch	cow	cup	daisy	dog
donut	elephant	fork	garlic	giraffe
grapes	handbag	horse	hydrangea	iris
kale	keyboard	kite	knife	laptop
lily	lime	mango	microwave	motorcycle
onion	orchid	oven	peony	pineapple
pizza	pomegranate	refrigerator	remote	rose
sandwich	sheep	sink	skateboard	skis
snowboard	spoon	strawberry	suitcase	sunflower
surfboard	tie	toaster	toilet	tomato
toothbrush	train	truck	tulip	tv
umbrella	vase	watermelon	zebra	

Table 3: List of 84 common entities used to study models' ability to generate multiple entities.

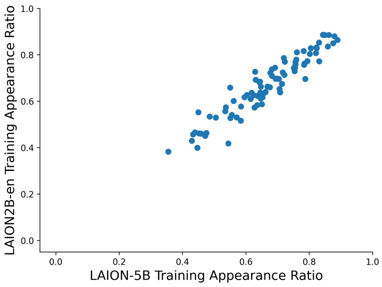
# Entities	Prompt
1	1. a/an [E1] 2. a photo of a/an [E1]
2	1. a/an [E1] and a/an [E2] 2. a photo of a/an [E1] and a/an [E2] 3. [E1], [E2] 4. a/an [E1] next to a/an [E2]
3	1. a/an [E1] and a/an [E2] and a/an [E3] 2. a photo of a/an [E1] and a/an [E2] and a/an [E3] 3. [E1], [E2], [E3] 4. a/an [E1] next to a/an [E2] and a/an [E3]

Table 4: Image generation prompts for single, double, and triple entities. [E1], [E2], and [E3] are replaced with various entities (e.g., elephant, zebra, and giraffe).

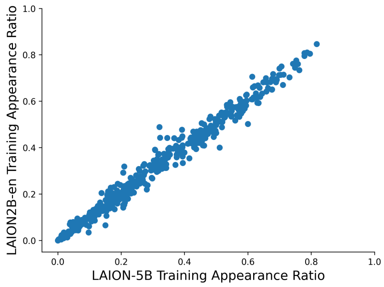
authors of this paper labeled 400 randomly selected generated images in the two entity setting, providing annotations for `entity1` and `entity2`. We find that PaliGemma predictions match human annotations in 90.88% of cases, which indicates strong performance. The biggest disagreements between human annotations and model predictions tend to be cases for which entities are similar in appearance and use cases (e.g., backpack and handbag), as well as large size differences (e.g., toothbrush and snowboard).

A.6 Comparing Training Appearance Ratios

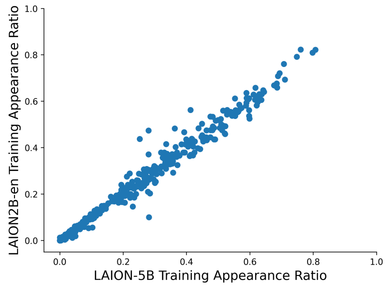
As shown in Figure 3, training appearance ratios calculated using LAION2B-en and LAION5B are highly correlated. While this is perhaps not surprising given that we focus exclusively on English and LAION2B-en is a subset of LAION5B, it is worth noting that these ratios are preserved across both datasets for the entity combinations we consider.



(a) Single Entity ($\rho = 0.95$)



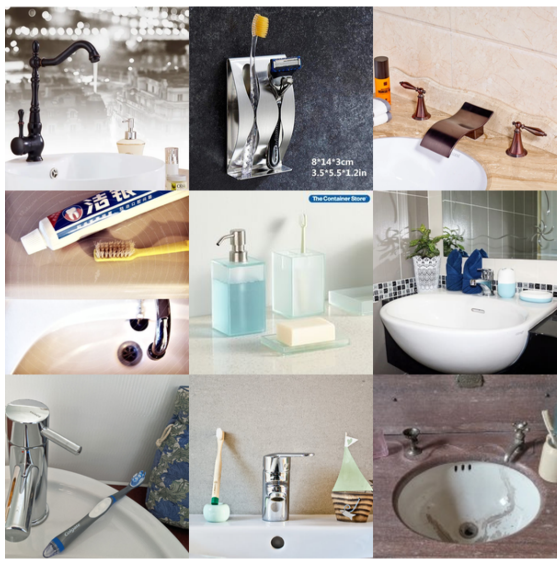
(b) Double Entities ($\rho = 0.99$)



(c) Triple Entities ($\rho = 0.99$)

Figure 3: Correlations between training appearance ratios ($tar_{e,ic}$) for LAION2B-en and LAION-5B for 1, 2, and 3 entities. We observe strong correlations for all three.

(a) Generated images using SD2.1 with the prompt "a toothbrush and a sink" ($gar_{e,ip}$=0.44).



(b) Training images whose captions mention both "sink" and "toothbrush" ($tar_{e,ic}$=0.44).

Figure 4: We sample generated and training images for the prompt "a toothbrush and a sink". Both the generation and training appearance ratios are the same. We see that generated images depicting one entity tend to show sinks, while training images depicting one entity show both toothbrush and sink individually.

(a) Generated images using SD2.1 with the prompt “a watermelon and a handbag” ($gar_{e,ip}$=0.48).

(b) Training images whose captions mention both “watermelon” and “handbag” ($tar_{e,ic}$=0.46).

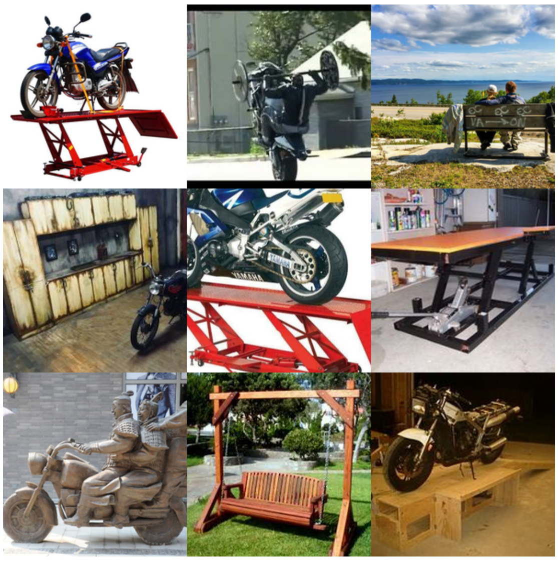
Figure 5: We sample generated and training images for the prompt “a watermelon and a handbag”. Both the generation and training appearance ratios are very similar. We see that generated images seem to always depict watermelons, and sometimes handbags (with appearances similar to a watermelon). While some training images are watermelon handbags, other examples may depict accessories or watermelon-colored handbags.

(a) Generated images using SD2.1 with the prompt “a giraffe and a bear” ($gar_{e,ip}$=0.46).

(b) Training images whose captions mention both “giraffe” and “bear” ($tar_{e,ic}$=0.43).

Figure 6: We sample generated and training images for the prompt “a giraffe and a bear”. We observe that while the generation and training appearance ratios are highly similar, the ways in which entities are depicted at generation and training differ quite noticeably (e.g., training images mostly show toys or cartoons).

(a) Generated images using SD1.5 with the prompt "a motorcycle and a bench" ($gar_{e,ip}$=0.08).



(b) Training images whose captions mention both "motorcycle" and "bench" ($tar_{e,ic}$=0.08).

Figure 7: We sample generated and training images for the prompt "a motorcycle and a bench". The generation and training appearance ratios are identical. At generation, the model generates images of motorcycles individually a clear majority of the time. The training data, however, also includes images of benches individually as well as images without either entity.

(a) Generated images using SD1.5 with the prompt "a photo of a bus and a horse" ($gar_{e,ip}$=0.18).

(b) Training images whose captions mention both "motorcycle" and "bench" ($tar_{e,ic}$=0.21).

Figure 8: We sample generated and training images for the prompt "a photo of a bus and a horse". The generation and training appearance ratios are very close. At generation, the model often generates buses individually, specifically red buses. While training images also depict buses individually in several cases, they seem to capture a more diverse set of buses.

(a) Generated images using SD1.5 with the prompt "elephant, daisy" ($gar_{e,ip}$=0.24).

(b) Training images whose captions mention both "elephant" and "daisy" ($tar_{e,ic}$=0.30).

Figure 9: We sample generated and training images for the prompt "elephant, daisy". The generation and training appearance ratios are fairly close. At generation, the model mostly depicts elephants individually, and they look reasonably realistic. In training images, we mainly see artistic renditions of elephants.

(a) Generated images using SD1.5 with the prompt "boat, chair" ($gar_{e,ip}$=0.16).

(b) Training images whose captions mention both "boat" and "chair" ($tar_{e,ic}$=0.19).

Figure 10: We sample generated and training images for the prompt "boat, chair". The generation and training appearance ratios are fairly close. At generation, the model primarily depicts a boat or chair, often individually, in an outdoor setting. In training images, while we see some entities in outdoor setting, many just depict a chair in a staged setting.

Image	Caption	VQA Predictions
	How To Make An **Asparagus Bed**	asparagus: yes, bed: no
	Bluetooth Speaker **Panda** with **Remote** Shutter Release White 4.3x4.5cm	panda: yes, remote: no
	Candy **apple** Red Volkswagen **bus** for couple and bridal party at water-front wedding	apple: no, bus: yes
	Sweet potato, **coconut** and **tomato** lentil dahl in a bowl beside a bowl of cherry tomatoes	coconut: no, tomato: yes
	Extreme BMX **Bicycle** Riding in Concrete **Skateboard** Park - Bar spin to tire tap Stock Footage	bicycle: yes, skateboard: no
	**Lily** the Borzoi chasing other **dog**	lily: no, dog: yes
	LED Waterproof RGB Colorful Wedding Party **Vase** Base Light Sub-mersible+**Remote**	vase: no, remote: yes
	An **elephant cow** taking a dust bath with her calf (Kruger National Park, South Africa).	elephant: yes, cow: no
	Collapsible **Chair** From **Skis** Ski Woodcraft Pinterest	chair: yes, skis: no
	Jungle Animal Shapes - **Cake** Toppers or Party Decorations monkey **giraffe** lion elephant tiger zebra snake hippo baby shower birthday party	cake: no, giraffe: yes

Table 5: Example training images and captions for which captions mention two specified entities (captions may mention other entities as well), but images only depict one of the specified entities clearly. Specified entities are in **bold**. One potential explanation for such occurrences is the ambiguity of words (e.g., "Lily" is both a name and a flower). Another explanation is that a combination of entities may have their own meaning (e.g., "asparagus bed" is not the same as "asparagus" + "bed").