Skews in the Phenomenon Space Hinder Generalization in Text-to-Image Generation

Yingshan Chang[1] Yasi Zhang[2] Zhiyuan Fang[3]
Ying Nian Wu[2] Yonatan Bisk[1] Feng Gao[3]

[1] Carnegie Mellon University yingshac@andrew.cmu.edu***
[2] University of California, Los Angeles
[3] Amazon†

Abstract. The literature on text-to-image generation is plagued by issues of faithfully composing entities with relations. But there lacks a formal understanding of how entity-relation compositions can be effectively learned. Moreover, the underlying phenomenon space that meaningfully reflects the problem structure is not well-defined, leading to an arms race for larger quantities of data in the hope that generalization emerges out of large-scale pretraining. We hypothesize that the underlying phenomenological coverage has not been proportionally scaled up, leading to a skew of the presented phenomenon which harms generalization. We introduce statistical metrics that quantify both the linguistic and visual skew of a dataset for relational learning, and show that generalization failures of text-to-image generation are a direct result of incomplete or unbalanced phenomenological coverage. We first perform experiments in a synthetic domain and demonstrate that systematically controlled metrics are strongly predictive of generalization performance. Then we move to natural images and show that simple distribution perturbations in light of our theories boost generalization without enlarging the absolute data size. This work informs an important direction towards quality-enhancing the data diversity or balance orthogonal to scaling up the absolute size. Our discussions point out important open questions on 1) Evaluation of generated entity-relation compositions, and 2) Better models for reasoning with abstract relations.

Keywords: Text-to-Image · Generalization · Relational Learning

1 Introduction

A visual scene is compositional in nature [51]. Atomic concepts, such as entity and texture, are composed via relations [18]. Relations represent abstract functions that are not visually presented, but modulate the visual realization of concepts. For example, consider a scene "*a cat is chasing a mouse*". It consists of atomic concepts: cat and mouse. "Chasing" defines a relation that is visually realized as certain postures and orientations of the cat and the mouse. Relations take concepts to fill their roles as functions take variables to fill their arguments. This process is known as role-filler binding [2,12,14,41], where fillers are concrete values and roles are abstract positions.

*** https://github.com/zdxdsw/skewed_relations_T2I

† This work is not related to these authors' position at Amazon.

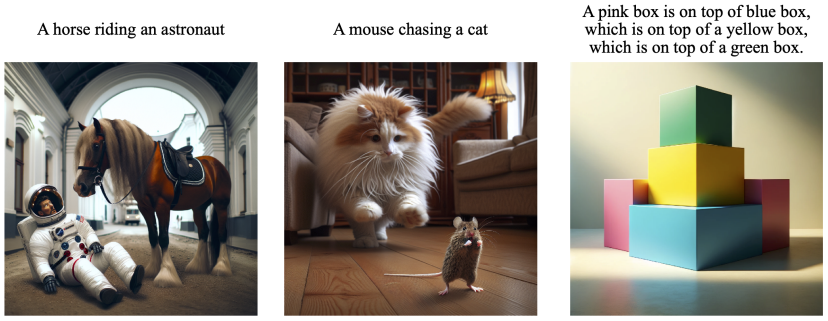


Fig. 1: Example images generated by DALL·E3. In all three cases, entities and relations are common but their compositions are uncommon. DALL·E3 tends to (a) compose entities unnaturally, (b) get trapped by the canonical relation, or (c) disregard the requested ordering. These errors are recurring across multiple trials, suggesting that DALL·E3 does not grasp the abstract notion of relations.

Due to abstractness, relations are always bound to concrete concepts in the space of observations, posting the challenge of truly grasping the abstract function of a relation and using it in generalizable ways, i.e. composing familiar relations with novel concepts. Recently, pre-trained text-to-image models [3,4,40] unleash the power of image synthesis with unprecedented fidelity and controllability. However, as shown by Figure 1, a pre-trained model cannot generate images faithful to the relational constraints upon seeing uncommon entity-relation compositions. This implies that, pre-trained text-to-image models do not represent role-filler bindings independently of the fillers, leading to an important question of **what hinders the learning of generalizable relations**.

This work investigates this question from the data distribution angle. We conjecture that although pre-training ensures massive data quantity, it does not accomplish a proportionally large coverage of unique phenomena. Figure 2 shows our conceptual framework for text-to-image generation, consisting of three distinct components: A text encoder, a visual decoder and a mechanism to communicate between these two spaces. We formalize the underlying structure of the data as role-filler bindings which nicely capture the compositional connections between data points. We assume that architectural expressivity and pretraining already enable both the text encoder and the visual decoder to distinctly represent fillers and roles in their corresponding spaces. Based on this assumption, the communication channel becomes the key to task success. We believe the choice of supervision data crucially affects the behavior of the communication channel.

To this end, we introduce two metrics that quantify the skew of the underlying structure supported by a dataset. These two metrics take into account linguistic notion of roles and visual notion of roles respectively. Our hypothesis is that generalization failure of text-to-image model is a direct result of phenomenological incompleteness or imbalance under our metric. Our experiments in both synthetic images and natural images demonstrate the strong predictive power of our metrics on generalization performance. We also show that findings from pixel diffusion models carry over to latent diffusion models.

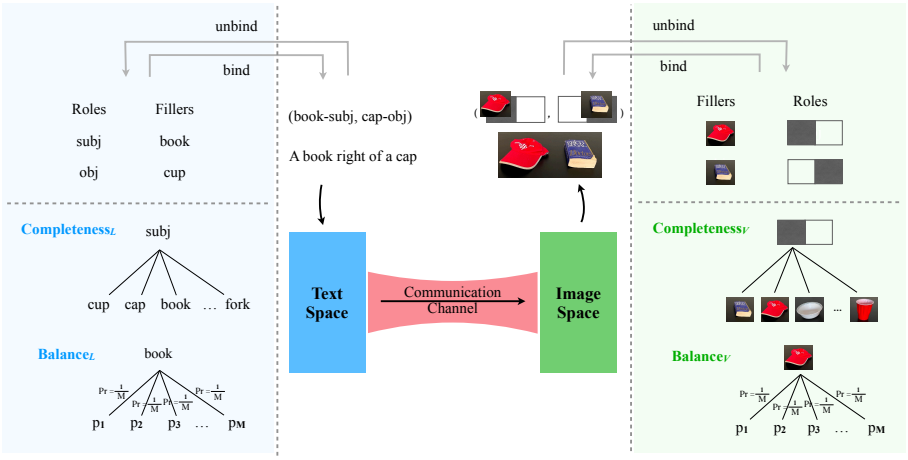


Fig. 2: Conceptual Framework. Text-to-Image generation consists of three important distinct components: A text encoder, a visual decoder, and a mechanism to communicate between these two spaces. Generation of images with consistent spatial relations requires that 1) the text encoder distinctly encodes linguistic roles, 2) the image generator distinguishes spatial roles in the output space, and 3) learning the correct translation from linguistic roles to visual roles. Suppose pre-training or architectural expressivity can fulfill the first two requirements, the remaining core task is to learn an effective communication channel – often instantiated as cross-attention in diffusion models. To this end, we propose statistical metrics to formally quantify how the training data distribution received by the communication channel affects generalization.

2 Related Work

Text Conditioned Image synthesis Diffusion models initiate the tide of synthesizing photorealistic images in the wild. They benefit from training stability and do not exhibit mode collapse that GAN models suffer from. [50] feeds text prompts to the diffusion model to make the generation process controllable. Inspired by ControlNet, a myriad of works [21, 28, 35, 40, 42, 47] explore the integration of text encoders and image generators, such that image synthesis, editing and style-translation can be customized by users via text. Unlike diffusion models, Transformer-based image synthesis models are naturally better at working in coordination with text, due to the shared tokenization process [4, 23]. Transformer-based approaches perform on par with diffusion on fidelity, and are believed to have greater potential for resolving long-range dependency and relational reasoning [30], thanks to their patch-based representations and attention blocks. However, Transformers suffer from a discrete latent space and slow inference speed. The latest work [30] integrates a Transformer architecture and diffusion objectives, aiming at the best of both worlds.

Despite high fidelity scores, a recurring problem is the difficulty to generate objects in unfamiliar relations [25]. Although those unfamiliar relations rarely occur in a natural collection of images, they are not physically implausible, and

humans have no trouble producing a corresponding scene. This has drawn attention to evaluating generative models and characterizing such failures. Works along this vein suggest that generative models typically fail at multiple objects, with multiple attributes or relations [8,13], where generalization to novel combinations of familiar components is needed the most.

Compositional generalization in image synthesis Compositional generalization is a specific form of generalization where individually known components are utilized to generate novel combinations. This remarkable learning ability has been widely studied in the vision-and-language understanding domain [15,19,22,32,38,39,44]. The text-to-image literature has recently seen efforts towards constructing images compositionally [7,45,49]. Closest to ours are two previous works that characterize properties of the underlying phenomenon space not trivially revealed by the pixel space. [29] has investigated shape, color and size as domains of atomic components, which can be combined to form tuples, e.g. (big, red, triangle). They mainly argue that generalization occurs under two conditions: 1) small structural distance between training and testing instances, and 2) effectively learning the disentanglement of attributes (i.e. a change in the size input will not affect the color output). [43] first assumes compositional data are formed by combining individually complex components with simple aggregation functions. Then they defined *compositional support* and *sufficient support* over a set of components, which are sufficient conditions for a learning system that compositionally generalizes.

Motivated by failures of existing methods, we investigate data-related factors that affect the generalization performance. There is a possibility that better architectural design can complement high-quality data to achieve generalization.

3 Formalization

We start by formalizing scene construction as role-filler bindings. A scene is constructed by binding fillers denoted as $\mathbf{F} = (f_1, \ldots, f_K)$ to roles denoted as $\mathbf{R} = (r_1, \ldots, r_K)$. Roles and fillers are paired up by their indices. Hence, each scene representation involves the same number of roles and fillers, i.e., $|\mathbf{F}| = |\mathbf{R}| = K$. Using ψ to denote the *binding* operation, a scene can be formalized into: $\psi(\mathbf{F}, \mathbf{R}) = (f_1/r_1, ..., f_K/r_K)$ and we call (f_k, r_k) a role-filler pair.

We would *unbind* a filler from $\psi(\mathbf{F}, \mathbf{R})$ via the unbinding operation ψ^{-1}: $\psi^{-1}\Big(\psi(\mathbf{F}, \mathbf{R}), f_k\Big) = r_k$. The unbinding operation describes the process of extracting the role from a binding that a given filler has been bound to, which corresponds to the decomposition of a compositional structure. Assume that ψ^{-1} returns *null* if the input filler f_k does not exist in the scene.

Fillers are atomic concepts that can be selected from a set of concepts $\mathcal{C} = \{c_1, ..., c_N\}$ while roles can take values from a set of candidate positions $\mathcal{P} = \{p_1, \ldots, p_M\}$. Typically, roles can have intrinsic meanings independent of the meanings of fillers [41]. For instance, the meaning of “upper position" is determined by the y-axis in a 2D image coordinate system, which exists independently of the specific pixel values that fulfill this role in each image. The

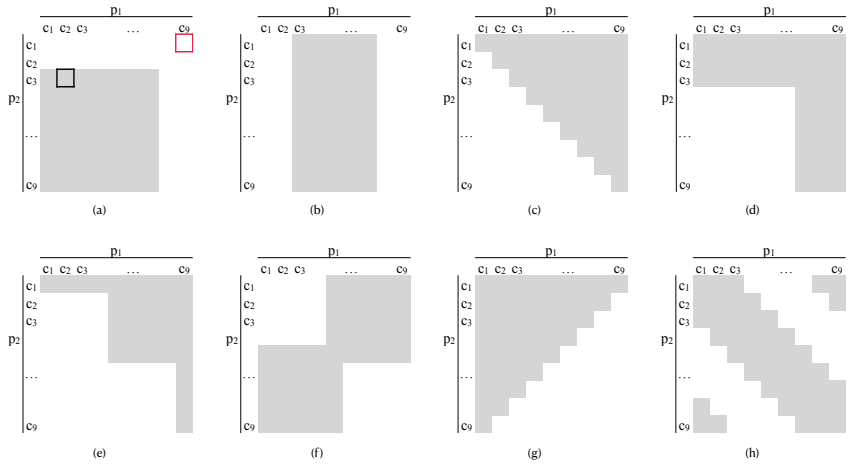


Fig. 3: Sketched illustrations of phenomenological coverage with different properties. Shaded areas represent the training set, while blank areas represent the testing set. Columns and rows are organized by the concepts bound to position 1 and position 2 respectively. For example, the black cell in (a) represents the training instance $(c_2/p_1, c_3/p_2)$, the red cell in (a) represents the testing instance $(c_9/p_1, c_1/p_2)$. (a) Both positions are incomplete (b) Only p_1 is incomplete (c) Complete but unbalanced (d) Complete but unbalanced (e) Complete but unbalanced (f) Complete and balanced (g) Complete and balanced (h) Complete and balanced

meanings of roles can be either learned from the task structure or manually designed. In the text-to-image case, the task structure naturally invites two ways to define roles, corresponding to the linguistic and the visual space, respectively. From the linguistic perspective, we consider grammatical positions, e.g. $\mathcal{P}_L = \{subject, object\}$. From the visual perspective, we consider spatial positions, e.g. $\mathcal{P}_V = \{top, bottom\}$.

Under our definitions, each image is a scene. Therefore, an image dataset can be essentially abstracted as a collection of bindings: $\mathcal{D} = \{\psi(\mathbf{F}^i, \mathbf{R}^i)\}_{i=1,\ldots,|\mathcal{D}|}$, where $\mathbf{F}^i$ and $\mathbf{R}^i$ denotes the fillers and roles in the i-th image. Let $\mathcal{U} = \mathcal{C} \times \mathcal{P}$ be the universe of all possible bindings. The vanilla notion of coverage supported by a dataset is the proportion of $\mathcal{U}$ that has non-zero supporting examples in the dataset: $\textbf{Coverage}(\mathcal{D}) = |Deduplicate(\mathcal{D})|/|\mathcal{U}|$, where the *Deduplicate* removes examples with equivalent role-filler representations. We argue that this metric overlooks how elements in $\mathcal{U}$ are structurally connected. For instance, each element in $\mathcal{U}$ shares a common role or filler with other elements. Without taking this structural property into account, truly meaningful coverage of diverse and unique phenomena might be conflated by the seemingly diverse surface forms.

Motivated by this consideration, our proposed metrics aim to measure whether a dataset has support for every concept occurring in every position, as well as the probability distribution of the positions that each concept has been bound to. Next, we formally describe completeness and balance metrics by conveniently leveraging the notations of binding and unbinding operators.

3.1 Completeness

Completeness requires that every relation has been bound with every concept across the entire dataset. In this sense, we define:

$$\textbf{Completeness}(p_m, \mathcal{D}) = \frac{\left|\left\{c_n \;\middle|\; p_m \in \psi^{-1}(\mathcal{D}, c_n), c_n \in \mathcal{C}\right\}\right|}{|\mathcal{C}|}, \tag{1}$$

where the operator extracting a concept from a dataset is defined as:

$$\psi^{-1}(\mathcal{D}, c_n) = \bigcup_{i=1}^{|\mathcal{D}|} \left\{\psi^{-1}\left(\psi(\mathbf{F}^i, \mathbf{R}^i), c_n\right)\right\}. \tag{2}$$

A fully complete dataset should have completeness scores of 1 for all relations. We take the expected value to obtain an aggregated score over the entire dataset:

$$\begin{aligned} \textbf{Completeness}(\mathcal{D}) &= \mathbb{E}\left[\textbf{Completeness}(p_m, \mathcal{D})\right] \\ &= \sum_{p_m \in \mathcal{P}} \mathbb{P}(p_m)\textbf{Completenss}(p_m, \mathcal{D}). \end{aligned} \tag{3}$$

3.2 Balance

Balance requires that every concept is bound with each position with equal probability. Concretely, we first calculate the entropy of all positions that a given concept c_n was bound to within dataset $\mathcal{D}$:

$$\textbf{Balance}(c_n, \mathcal{D}) = \mathit{Entropy}\left[\psi^{-1}\left(\psi(\mathbf{F}^i, \mathbf{R}^i), c_n\right), i = 1, ..., |\mathcal{D}|\right], \tag{4}$$

where the function *Entropy* computes the entropy of the distribution of the positions. Note that we constrain the computation to practical positions, i.e., excluding *null* during the calculation process. Then we aggregate by taking the expected value to obtain the balance over the entire dataset:

$$\textbf{Balance}(\mathcal{D}) = \mathbb{E}\left[\textbf{Balance}(c_n, \mathcal{D})\right] = \sum_{c_n \in \mathcal{C}} \mathbb{P}(c_n)\textbf{Balance}(c_n, \mathcal{D}). \tag{5}$$

This metric is upper bounded by $\log(M)$, corresponding to the entropy of a uniform distribution over $\mathcal{P}$. The lower the metric, the higher the skew of a dataset. We normalize by $\log(M)$ to obtain a value within $[0, 1]$.

So far we have defined completeness and balance for arbitrary sets of fillers and roles. Hereinafter, we will focus on binary relations, i.e. $|\mathbf{F}| = |\mathbf{R}| = 2$. Subscripts $_L$ and $_V$ denote the linguistic and visual perspectives under which a metric is computed. Practically, metrics are estimated by empirical counts. Next, we conduct experiments to demonstrate that generalization is hindered by incompleteness or imbalance under either perspective.

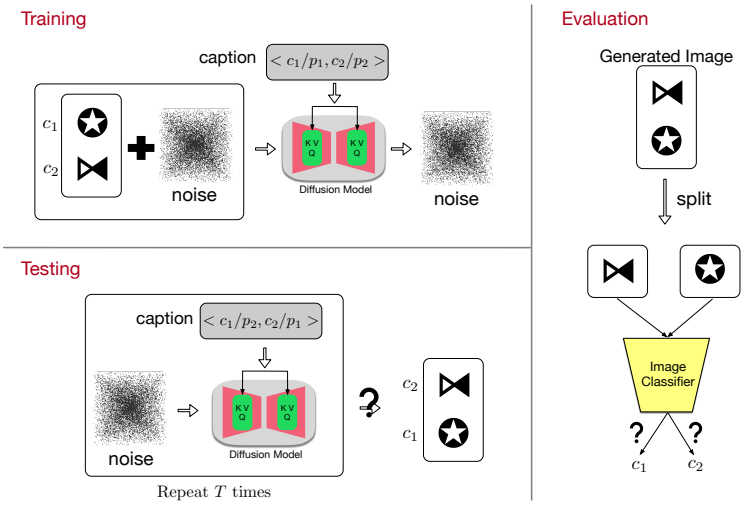


Fig. 4: Training, testing and evaluation pipeline. We train diffusion models to generate images of two concepts (c_1, c_2) with a specified spatial relation. Then the model is tested on unseen concept pairs to see whether the learned relations are generalizable.

4 Experiments on Synthetic Images

4.1 Setup

We use a set of unicode icons as concepts, and assign common nouns to them as their names. We vary the number of concepts, N, within $\{30, 40, 50, 60, 70, 80, 90\}$. We consider two symmetric binary relations: "on top of", "at the bottom of". Images are constructed by drawing each icon on a 32x32 canvas, then stacking two icons vertically, resulting in 32x64 resolution. Captions are created from the template: "a(an) <icon_name> is <relation> a(an) <icon_name>".

Training sets ($\mathcal{D}$) are sampled from $\mathcal{U}$ with systematic control for the four properties: $\textbf{Completeness}_L$, $\textbf{Completeness}_V$, $\textbf{Balance}_L$, $\textbf{Balance}_V$. To avoid confounders, we ensure the linguistic metrics are perfect when studying the effect of the visual metrics, and vice versa. Figure 3 illustrates training distribution with varied properties. We take the complementary set, $\mathcal{U} \setminus \mathcal{D}$ as the testing set. Testing instances are unseen in terms of the tuple representation (f_1, r_1, f_2, r_2). Yet we show that perfect testing accuracy is possible when the training set properly supports the phenomena space, i.e. being complete and balanced.

Our training, testing and evaluation pipeline is depicted in Figure 4. Following the architecture of [11], we train 350M UNet models on text-conditioned pixel-space diffusion, with T5-small [34] as the text encoder. The size of UNet is determined such that it is minimally above the threshold at which the model is capable of fully fitting the training set. This would avoid issues such as a lack of expressivity or under-training of an overparameterized model.[4] For evaluation,

[4] We adopt an lr of 1e-4 and batch size of 16. Evaluation is performed every 20 epochs. Early stopping is applied when the evaluation has not been better for 100 epochs.

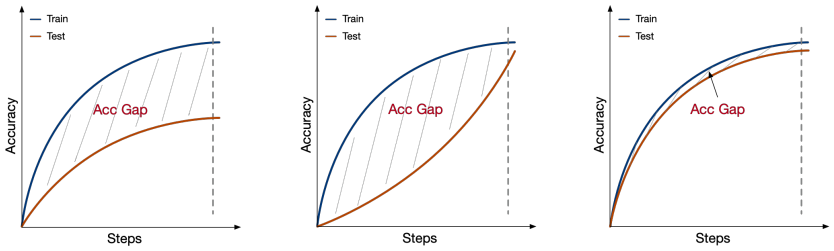


Fig. 5: Three types of learning dynamics are observed in our experiments. In the worst scenario (left), the testing accuracy plateaus and never converges. In the best scenario (right), the testing accuracy closely tracks the training accuracy until both converge to perfect. In the middle scenario (center), the testing accuracy climbs slower than the training accuracy, but is still able to converge to perfect at a delayed point after the training accuracy has already converged. In order to distinguish between the middle and best scenarios, we additionally report the accumulative gap between training and testing accuracy curves, which captures the timeliness of generalization.

we compute accuracy of both icons being generated correctly[5]. The evaluation process is automated by pattern-matching icons with convolutional kernels.

Alongside the final testing accuracy, we report the accumulative difference between training and testing accuracy curves. This is because we have observed a third type of learning dynamics lying in between a generalization success and generalization failure (Figure 5), where the testing accuracy climbs slower than the training accuracy, but is still able to converge to perfect at a delayed point Only reporting on final testing accuracy would mask the qualitative difference that captures a notion of how timely generalization occurs.

4.2 Results

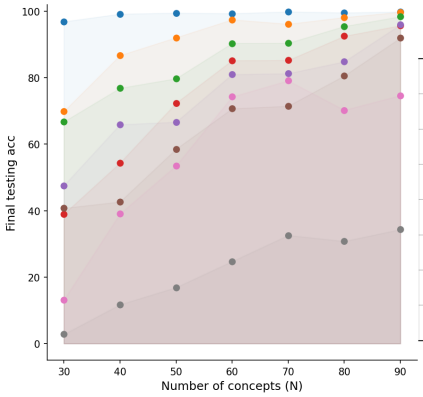
Visual incompleteness significantly impedes generalization. The last four experiments in Figure 6 (indexed by purple, brown, pink and grey) have low **Completeness**$_V$. Their testing performance never reached 100%, even plateauing below 50% for smaller concept sets.

Visual imbalance harms generalization when N is small. The first four experiments in Figure 6 (indexed by blue, orange, green and red) show the progression of increasing **Balance**$_V$ while keeping full **Completeness**$_V$. As **Balance**$_V$ grows, the testing accuracy consistently improves for all N. Increasing N can provide a remedy for a dataset with complete but imbalanced support. In contrast, larger N does not bring much help the support is incomplete.

Linguistic incompleteness or imbalance harms generalization to a lesser degree, but they delay generalization. Figure 7 (left) shows that all cases

The length of training typically falls between 600 epochs ($N = 30$) and 200 epochs ($N = 90$). A full list of model and training configs is provided in the appendix.

[5] During evaluation, we find that the random state at which each diffusion process begins with has negligible effect ($\pm 1\%$) on final performance



	CPL(r1)	CPL(r2)	BLC	Coverage
●	100	100	100	20
●	100	100	84	55
●	100	100	79	43
●	100	100	71	50
●	84	84	79	70
●	75	75	68	56
●	100	50	68	50
●	50	50	0	25

Fig. 6: Under the definition of roles (r_1, r_2) from **visual** perspective ("top", "bottom"), the plot (left) shows final testing accuracy against distributional properties of the training set. The legend and the corresponding metrics are summarized in the table (right). The results suggest that both **Completeness**$_V$ (CPL) and **Balance**$_V$ (BLC) are positively correlated with testing (generalization) performance. By contrast, a vanilla notion of data coverage is badly correlated with performance.

achieve perfect or near-perfect testing accuracy, unless for the very small concept classes. This suggests that linguistic incompleteness and imbalance do not severely hinder whether the model is able to generalize ultimately. However, they do bring a negative effect by delaying the onset of generalization. This delay effect is revealed in Figure 7 (right). The takeaway is that, although the final testing accuracy is comparable, lack of **Completeness**$_L$ or **Balance**$_L$ causes the testing acc to largely lag behind training acc, which takes a longer time to catch up. Similarly, we plot the generalization gap for the set of visual skew experiments in the appendix, observing the same trend. However, since both failing to generalize or having a prolonged duration before generalizing can lead to a large gap, this result has to be taken with a grain of salt.

Vanilla notions of coverage are a bad predictor for generalization. The rightmost columns in Figures 6 and 7 provide the training set's coverage (%) over the universe $\mathcal{U}$. It can be seen that this does not correlate with the generalization performance. For example, the green run in Figure 6 outperformed red, purple, brown and pink runs, while having a much lower coverage than them. Also noteworthy is that we intentionally select low-coverage datasets to demonstrate the fully-complete, fully-balanced case, which achieves perfect generalization for all N. This strongly indicates the problem of aiming for a generalizable model by recklessly scaling up coverage. In accordance with research on model bias, we argue that scaling up along the incorrect axes is dangerous because it may aggravate unintended bias, without necessarily benefiting generalization.

Increasing N eases generalization in all cases. The model consistently generalizes better when trained on more concepts, albeit to varying degrees. Figures 6 suggests that enlarging N is more helpful when the data is within a decent range of **Completeness**$_V$ and **Balance**$_V$ (e.g. 70-80).

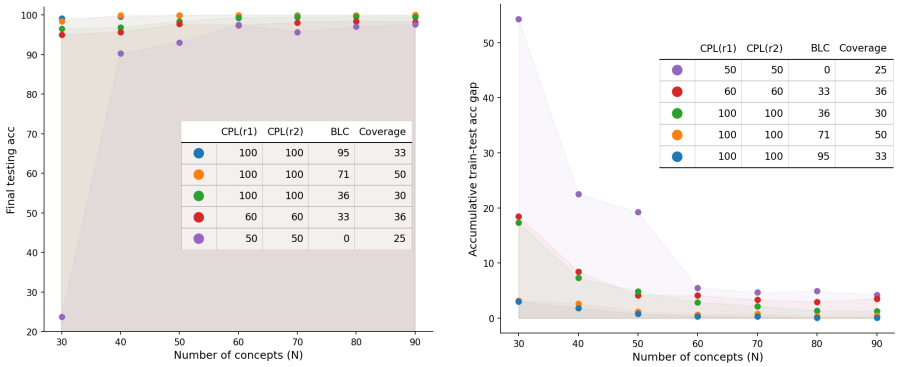


Fig. 7: Under the definition of roles (r_1, r_2), **linguistic**ally ("subj", "obj"), the plot shows final testing accuracy (left) and train-test accuracy gaps(right) against distributional properties of the training set. Legend and metrics are summarized in the table. The left plot suggests that linguistic incompleteness and imbalance harm generalization for small concept classes, while having less impact on the final testing accuracy for large concept classes. The right plot suggests that linguistic incompleteness and imbalance harm the timeliness of generalization, as indicated by larger train-test accuracy gaps.

5 Experiments on Natural Images

5.1 Setup

We extend our experiments to natural images using the What'sUp benchmark proposed by [16].[6] Our hypothesis is that higher Completeness and Balance of the training distribution lead to more generalizable learning outcomes. Each image in the What'sUp benchmark contains two objects, with a caption describing their spatial relations. The spatial relations include two pairs of symmetric binary relations: *left/right of* and *in-front/behind*. Without loss of generality, we study *left/right of*, leaving the exploration of more than two relations for the future.

From an initial (complete and balanced) set of 308 samples with 15 unique concepts, subsamples are drawn where completeness and balance vary. For fair comparisons, the coverage of all subsamples is relatively the same. We train 470M pixel-space diffusion models on What'sUp image-caption pairs, with 64x32 resolution, 5e-4 learning rate and a batch size of 16. Early stopping is applied similarly to the synthetic setting. The length of training typically falls between 3000 and 6000 epochs. Hyperparameter tuning is described in the appendix.

Since under the natural image setting the same object may occur at different image positions, evaluation could not be performed with predetermined pattern-matching kernels. We finetune ViT-B/16 [6] as an automatic evaluation engine to classify the object in the left or right crop of generated images.[7] A

[6] The What'sUp benchmark provided subsetA and subsetB. We adopt subsetB for our purpose because objects in subsetA have a great disparity in their sizes.

[7] The auto-eval engine can be deemed oracle. On 144 manually checked samples, the ViT's judgments are all correct

Table 1: What'sUp benchmark results varying **Completeness**$_V$ and **Balance**$_V$

Training Set Properties				Performance	
CPL(r_1)	CPL(r_2)	BLC	Cov.	Final Acc$^\uparrow$	Acc Gap$^\downarrow$
100	50	63	47	18.75	84.55
80	73	77	50	19.52	73.87
87	87	75	49	25.93	68.41
100	100	73	44	10.17	80.49
100	100	88	49	28.50	64.89
100	100	100	48	48.64	33.67

Table 2: What'sUp benchmark results varying **Completeness**$_L$ and **Balance**$_L$

Training Set Properties				Performance	
CPL(r_1)	CPL(r_2)	BLC	Cov.	Final Acc$^\uparrow$	Acc Gap$^\downarrow$
50	100	63	47	57.14	40.72
80	73	77	50	60.00	39.83
87	87	75	49	62.04	38.44
100	100	73	44	62.71	33.90
100	100	80	37	65.04	30.21
100	100	88	49	67.29	32.90

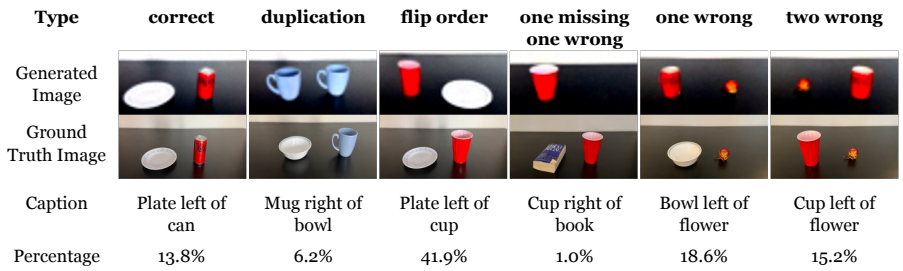


Fig. 8: Qualitative examples of generated images and the corresponding ground truth images, including a categorization of failure types and their frequencies.

"blank" label is added to the classifier in order to indicate when a model fails to generate an object. Importantly, a secondary result of our work is demonstrating shortcomings of the widely used evaluation methods for image synthesis, such as CLIPScore [10], VQA with VLMs [13,26], bounding-box evaluation with detectors [8,13]. See Section 6 for discussions of when they fall short.

5.2 Results

On natural images, it is much harder to generalize, probably because the objects' absolute position and postures can vary even when the relative spatial positions are determined. Another likely cause is the small number of concepts, as suggested by Section 4.2 that generalization tends to plateau at 50% for a small concept set. Nevertheless, the relative performance across different training set properties still conveys a meaningful message. Table 1 shows a consistent trend of generalization performance influenced by **Completeness**$_V$ and **Balance**$_V$. The final accuracy gets higher and the generalization gap gets lower as the training distribution moves towards being fully complete and balanced. **Completeness**$_L$ and **Balance**$_L$ achieve a similar effect, as shown by Table 2.

As with our previous findings, visual skew imposes a more severe challenge, compared with the same amount of distributional skew on the linguistic side. Our explanation is that, since the network is modeling the pixel space, it is more

directly affected by the skew of the observed pixel-space distribution, while skew of the language-space distribution impacts the result rather indirectly.

Upon closer examination of model outputs, the most common error was generating the correct objects with flipped order. This suggests that mapping fillers across domains is easy, but learning to map roles when only observing role-filler bindings is hard. Other errors include generating a blank or duplicating the object. Figure 8 visualizes examples of correct and incorrect generations.

6 Discussions

We present a conceptual framework and formal metrics to study the contributing factors of generating images with correct spatial relations. Clearly, our work triggers many open questions that are worth future exploration. Appendix I further discusses limitations of this work and future directions.

Are spatial relations distinguishable within unimodal spaces? We have mainly focused on what enables entity-relation compositions to be successfully conveyed from the text to the image domain. However, this question is meaningful only under the assumption that different roles and fillers are distinctly encoded in unimodal spaces. We have evidence that, perhaps surprisingly, this assumption does not always hold in existing approaches.

We train probing classifiers to extract the positional role of nouns from text encodings. More details are available in Appendix E. We find that probes trained with the CLIP text encoder can only overfit the training data, but not generalizing. This indicates the inherent weakness of CLIP text encoder to provide consistent signals of spatial positions. This finding aligns with existing criticism on CLIP essentially being a bag-of-word model [48]. By contrast, probing experiments with T5 [34] encoder and the encoder of a pretrained vision-language model [9] succeed with near-perfect generalization. This makes T5 and VLMs naturally better candidates for training text-to-image models.

On the image side, the capacity to represent positions can be theoretically guaranteed by image-patch positional encodings, readily compatible with attention blocks in the diffusion architecture. However, to our surprise, most of the open-source diffusion implementations [31] omit this step. We noticed this issue as our initial experiments failed unexpectedly. The problem was fixed after we modified the architecture to include image positional encodings. We posit that the lack of image positional encodings imposes a representational deficiency[8] leading to heavy reliance on pixel correlations and unexpected testing behavior. Ablation studies in Appendix F support this argument.

In short, we point out the importance of a text encoder that distinguishes positional roles and an image decoder that has the representational power for spatial information. We emphasize that these are not only important preconditions for our main analysis presented in this paper, but should also be crucial considerations in future generative models or models that do spatial reasoning.

[8] The zero paddings in convolutional layers can possibly leak positional information, but they need many layers to propagate information from the periphery to the center.

Generation in the Latent Space There are abundant studies on latent generation methods [4, 30, 36, 37] that are able to achieve higher resolution. However, the progress towards spatially consistent high-resolution synthesis might be hampered when the data coverage of underlying phenomena does not proportionally scale with resolution. To test whether our arguments carry over to latent-space generation, we conduct experiments with a pretrained (frozen) VAE (stabilityai/stable-diffusion-2-1). We increase the input resolution by a factor of four because VAE applies compression. The results (Appendix G) match our intuition that a latent space does not affect the validity of our conceptual and formal frameworks. CPL and BLC consistently correlate with generalization performance. Also in accordance with our existing findings, linguistic skew harms final testing performance to a lesser degree, but delays generalization, as evidenced by large acc gaps. Finally, the latent diffusion results again verified that larger coverage cannot compensate for a poor CPL or BLC.

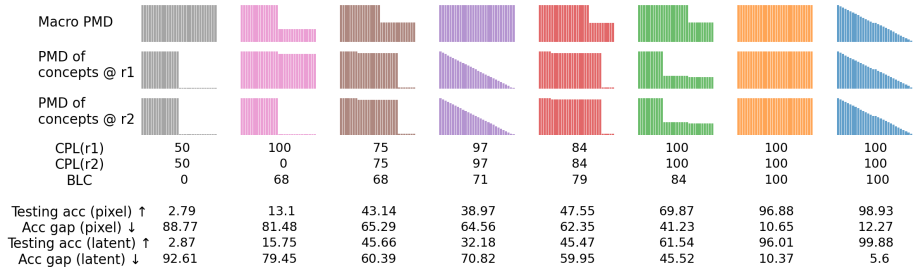
Two factors may explain the similarity in results between pixel and latent spaces. First, the latent space feature maps have spatial correspondence with the image. Second, the VAE [17] does not have a language component, as such the language-to-vision communication channel is captured by diffusion. While better features may be provided by VAE, they lack any cross-modal correspondence. In summary, increasing the phenomenological coverage and increasing resolution are both important. We leave the questions open on extending our formal notions to superresolution models as well as more nuanced relations.

Text-to-Image Evaluation Methods We rely on several heuristics when automating the evaluation of generated images. This was feasible only when 1) objects are center aligned, and 2) the background is clean. Ultimately, we are interested in evaluating relations with cluttered scenes and with greater appearance variability. Besides finetuning a ViT classifier, we have explored existing off-the-shelf models, but found them limited in one way or another. CLIPScore is known to pay less attention to token orders. Indeed, CLIPScore judges correctly in only 37% of the times in our setting. Evaluating spatial relations by comparing bounding box locations produced by detectors offers a more structured approach, yet it is restricted to the object classes available in the detection pretraining. For example, “headphones" and “tape" are classes in the What’sUp benchmark that do not belong to any of the popular detection datasets, rendering most of the detection models inappropriate for our experiments. Open-vocabulary detectors [27] circumvents this problem. But practical issues still exist, such as redundancy in the predicted bounding boxes and sensitivity to text prompts.

Aside from using CLIP or detectors, the literature has also suggested using vision-language foundation models (VLMs) for synthesized images evaluation. In addition to the apparent drawback of slow inference, it remains questionable whether VLMs comprehend relations in the first place [48]. Our attempts at LLaVA- or BLIP2-VQA were unsuccessful for multiple reasons such as inability to recognize spatial relation (Table A3), and unbalanced precisions across object classes (Table A5). See Appendix H for more analysis. We hope our investigation

calls into question the effectiveness of existing metrics on spatial consistency, which might inadvertently mask the weakness of text-to-image models.

Other data distributional properties Our metrics are designed around completeness and balance, but they may not capture other distributional properties, such as the Zipfian-Uniform axis. Figure 9 plots the probability mass distribution of datasets we used in Section 4. We see three shapes: uniform, 2-stages, and "wedge". This provides new insights: 1) Macro-PMD may hide skew, visible when we plot role-specific-PMD (row 2&3). E.g. grey and purple instances (col 1&4) exhibit uniform macro-PMD, but biased role-specific-PMD – correlated to poor generalization. 2) A non-uniform PMD may not indicate poor generalization. E.g. the blue instance (col 8) has non-uniform PMD, yet achieves great performance – perfect CPL and BLC scores. Future work may explore other data distributional properties and their correlation with generalization.



CPL(r1)	50	100	75	97	84	100	100	100
CPL(r2)	50	0	75	97	84	100	100	100
BLC	0	68	68	71	79	84	100	100
Testing acc (pixel) ↑	2.79	13.1	43.14	38.97	47.55	69.87	96.88	98.93
Acc gap (pixel) ↓	88.77	81.48	65.29	64.56	62.35	41.23	10.65	12.27
Testing acc (latent) ↑	2.87	15.75	45.66	32.18	45.47	61.54	96.01	99.88
Acc gap (latent) ↓	92.61	79.45	60.39	70.82	59.95	45.52	10.37	5.6

Fig. 9: Probability mass distribution (PMD) of various training sets used in Section 4. Testing performance is correlated more with CPL and BLC, than with the uniformity of PMD. Also note that the PMD of concepts at individual roles may not be uniform although it appears to be uniform on the macro level.

7 Conclusion

Text-to-Image synthesis, despite recent breakthroughs in fidelity, appearance diversity and texture granularity, still struggles with relations. As the community strives for larger datasets to better cover the natural distribution, there is a lack of study on the axes along which phenomenological coverage can be meaningfully enlarged. This work presents the first effort to formally characterize training coverage, in the context of learning spatial relations. We introduce completeness and balance metrics under both the linguistic and visual perspectives. Our experiments on synthetic and natural data consistently suggest that models trained on more complete and balanced datasets have greater generalization potential. We see this work as a stepping stone towards text-to-image models that can faithfully generate relations in general, including implicit relations entailed by verbs. We hope to inspire research on structured image evaluation, architectures for modeling role-filler bindings and formal frameworks of generalization.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019)
2. Ajjanagadde, V., Shastri, L.: Rules and variables in neural nets. Neural Computation **3**(1), 121–134 (1991)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
4. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
5. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Goel, V., Peruzzo, E., Jiang, Y., Xu, D., Sebe, N., Darrell, T., Wang, Z., Shi, H.: Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. arXiv preprint arXiv:2303.17546 (2023)
8. Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)
9. Gui, L., Chang, Y., Huang, Q., Som, S., Hauptmann, A.G., Gao, J., Bisk, Y.: Training vision-language transformers from captions. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=xLnbSpozWS
10. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Holyoak, K.J.: Analogy and relational reasoning. The Oxford handbook of thinking and reasoning pp. 234–259 (2012)
13. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
14. Hummel, J.E., Holyoak, K.J., Green, C.B., Doumas, L.A., Devnich, D., Kittur, A., Kalar, D.J.: A solution to the binding problem for compositional connectionism. In: AAAI Technical Report (3). pp. 31–34 (2004)
15. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
16. Kamath, A., Hessel, J., Chang, K.W.: What's" up" with vision-language models? investigating their struggle with spatial reasoning. arXiv preprint arXiv:2310.19785 (2023)

17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
19. Lake, B., Baroni, M.: Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In: International conference on machine learning. pp. 2873–2882. PMLR (2018)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
21. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
22. Lindemann, M., Koller, A., Titov, I.: Compositional generalisation with structured reordering and fertility layers. arXiv preprint arXiv:2210.03183 (2022)
23. Liu, H., Yan, W., Abbeel, P.: Language quantized autoencoders: Towards unsupervised text-image alignment. Advances in Neural Information Processing Systems **36** (2024)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
25. Lovering, C., Pavlick, E.: Training priors predict text-to-image model performance. arXiv preprint arXiv:2306.01755 (2023)
26. Lu, Y., Yang, X., Li, X., Wang, X.E., Wang, W.Y.: Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. Advances in Neural Information Processing Systems **36** (2024)
27. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European Conference on Computer Vision. pp. 728–755. Springer (2022)
28. Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. arXiv preprint arXiv:2312.07536 (2023)
29. Okawa, M., Lubana, E.S., Dick, R., Tanaka, H.: Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. Advances in Neural Information Processing Systems **36** (2024)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
31. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
32. Potts, C.: Compositionality or generalization? (2019)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
34. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)

35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
38. Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., Lake, B.M.: A benchmark for systematic generalization in grounded language understanding. Advances in neural information processing systems **33**, 19861–19872 (2020)
39. Russin, J., Fernandez, R., Palangi, H., Rosen, E., Jojic, N., Smolensky, P., Gao, J.: Compositional processing emerges in neural networks solving math problems. In: CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference. vol. 2021, p. 1767. NIH Public Access (2021)
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
41. Smolensky, P.: Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial intelligence **46**(1-2), 159–216 (1990)
42. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
43. Wiedemer, T., Mayilvahanan, P., Bethge, M., Brendel, W.: Compositional generalization from first principles. Advances in Neural Information Processing Systems **36** (2024)
44. Wu, Z., Kreiss, E., Ong, D.C., Potts, C.: Reascan: Compositional reasoning in language grounding. arXiv preprint arXiv:2109.08994 (2021)
45. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023)
46. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (2014)
47. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
48. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: 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 (2022)
49. Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., Patel, V.M.: Scenecomposer: Any-level semantic image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22468–22478 (2023)
50. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
51. Zhou, Y., Feinman, R., Lake, B.M.: Compositional diversity in visual concept learning. Cognition **244**, 105711 (2024)