Learning to Model Multimodal Semantic Alignment for Story Visualization

Bowen Li
University of Oxford
bowen.li@cs.ox.ac.uk

Thomas Lukasiewicz
TU Wien, University of Oxford
thomas.lukasiewicz@cs.ox.ac.uk

Abstract

Story visualization aims to generate a sequence of images to narrate each sentence in a multi-sentence story, where the images should be realistic and keep global consistency across dynamic scenes and characters. Current works face the problem of semantic misalignment because of their fixed architecture and diversity of input modalities. To address this problem, we explore the semantic alignment between text and image representations by learning to match their semantic levels in the GAN-based generative model. More specifically, we introduce dynamic interactions according to learning to dynamically explore various semantic depths and fuse the different-modal information at a matched semantic level, which thus relieves the text-image semantic misalignment problem. Extensive experiments on different datasets demonstrate the improvements of our approach, neither using segmentation masks nor auxiliary captioning networks, on image quality and story consistency, compared with state-of-the-art methods.

1 Introduction

Story visualization is a challenging task, which aims to generate a sequence of story images given a multi-sentence story, and further requires output images to be consistent, e.g., having a consistent background or character appearance. Regardless of its difficulties, story visualization has the potential for many applications, including art creation, computer-aided design, and image editing.

To address the challenges, current methods (Li et al., 2019c; Song et al., 2020; Maharana et al., 2021; Maharana and Bansal, 2021) adopt a fixed StoryGAN-based (Li et al., 2019c) architecture, where two GANs (Goodfellow et al., 2014) are adopted, one for single-image quality, and one for story consistency, without considering the semantic alignment between different-modal text and image features involved in the generation process.

So, one problem arising is that a fixed network with the involvement of different-modal representations (e.g., text and image) may suffer from a semantic misalignment problem. This is because current methods usually adopt a fixed text encoder and image encoder to extract corresponding features, and then use these features directly in an also fixed GAN-based network. However, text representations can be a coarse sentence vector, fine-grained word embeddings, or a structured knowledge graph (Mahon et al., 2020), while Gatys et al. (2016) and Johnson et al. (2016) have shown that image features extracted from different layers of a convolutional neural network (e.g., VGG) may contain different-level semantic information. Based on this, simply fusing cross-domain representations together using a fixed generative network may cause a considerable adverse impact on the quality of output images. For example, one of the main function of the discriminator in a conditional GAN is to evaluate the semantic alignment between input text and output image, and provide the corresponding feedback to the generator, which encourages the generator to generate text-semantic-matched images. However, to evaluate the semantic alignment, the discriminator in current methods only simply concatenates a coarse sentence vector and image features at a small-scale (i.e., 4×4), extracted from a given image using a series of convolutional layers, which may fail to fully match the semantics between text and image features, and thus provide a less precise training feedback to the generator.

To address this problem, we explore the semantic alignment between text and image representations by learning to match their semantic levels in the GAN-based generative model. More specifically, we introduce dynamic interactions according to learning to dynamically explore various semantic depths and fuse the different-modal information at a matched semantic level. By doing this, the network can learn to dynamically utilize various

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4712–4718
December 7-11, 2022

semantic level information from the given representations, and also learn to selectively fuse them together, which thus mitigates the semantic misalignment problem across different modalities. So, the main contributions are summarized as follows:

• We fully explore the semantic misalignment problem existing in current methods, and propose a novel single-GAN based network, which improves FID from 78.64 to 52.87, and FSD from 94.53 to 55.20 on Pororo-SV, and establishes a benchmark FID of 74.12 and FSD of 20.07 on Abstract Scenes.

• We conduct extensive experiments and a thorough analysis of aligning the semantics between different-modal inputs to provide general modeling insights into conditional GANs.

2 Related Work

Story visualization aims to generate a sequence of consistent images corresponding to a multi-sentence story. StoryGAN (Li et al., 2019c) introduced a two-GAN-based generation network. CP-CSV (Song et al., 2020), DUCO (Maharana et al., 2021), and VLC (Maharana and Bansal, 2021) were built on StoryGAN, where CP-CSV utilized character segmentation masks to improve the performance, and DUCO and VLC adopted auxiliary captioning networks to build a text-image-text circle to ensure the consistency between the input and output. Recently, Li and Lukasiewicz (2022b) proposed to utilize fine-grained word information to build a concise single-GAN based network, and Li et al. (2022a) proposed to combine both clustering learning and contrastive learning together to ensure better text and image representations in a joint space. However, all these methods were based on a fixed network without considering the semantic alignment between involved text and image representations.

Text-to-image generation is closely related to our work, which generates one image from one given text description (Reed et al., 2016; Zhang et al., 2018; Qiao et al., 2019; Hinz et al., 2019; Tao et al., 2020; Zhu et al., 2019; Xu et al., 2018; Li et al., 2019a,b, 2020; Zhang et al., 2021; Li et al., 2022b). Differently, story visualization is more challenging, as it further requires output story images to be consistent.

3 Overview

Differently from current methods that adopt two GANs, our network only has a single GAN, neither requiring additional segmentation masks nor auxiliary networks for supervision, as we experimentally find that a single-GAN-based network can effectively produce high-quality story images with a good consistency. We attribute this improvement to the exploration of semantic alignment between different-modal representations, which enables fine-grained training feedback from the discriminator to the generator.

Given a story X with n story sentences, a text encoder encodes each story sentence S_i into a sentence vector $s \in \mathbb{R}^D$ with corresponding word embeddings $s_{\text{word}} \in \mathbb{R}^{D \times L}$, where D is the feature dimension, and L is the number of words in a sentence. Then, we feed this n sentence vectors into the generation pipeline using upsampling blocks to produce story images at the required resolution. Meanwhile, we further incorporate word embeddings into the generation pipeline, according to our proposed dynamic interactions, allowing the generator to learn to choose semantically aligned inputs from different semantic levels and modalities to achieve a better generation. In the discriminator, we also adopt the proposed dynamic interactions to fuse both text and image features, enabling a better evaluation on the text-image semantic alignment. The complete architecture is shown in Fig. 1.

3.1 Exploration of Semantic Alignment

To ensure the semantic alignment between text and image representations, we introduce dynamic interactions via utilizing self-attention and cross-modal attention, learning to dynamically explore various semantic depths from these different-modal representations, and also to fuse them together at a matched semantic level, which thus mitigates the text-image semantic misalignment problem.

3.1.1 Attention Mode

Two attention modes are adopted in our approach, one is self-attention (Zhang et al., 2019) (SA), and one is word-level spatial attention (Xu et al., 2018) (WSA). To generalize both types of attention, we set a as one input, and b as the other input, where a denotes intermediate image features in the generator or discriminator, and b denotes word embeddings in WSA, or the same image features as a in SA. So, the attention weights can be achieved via $\beta = \text{Softmax}(ab^T)$. Then, we can get weighted hidden features h via $h = \beta b$. Finally, we selectively fuse the weighted hidden features into the network using the proposed dynamic block (details

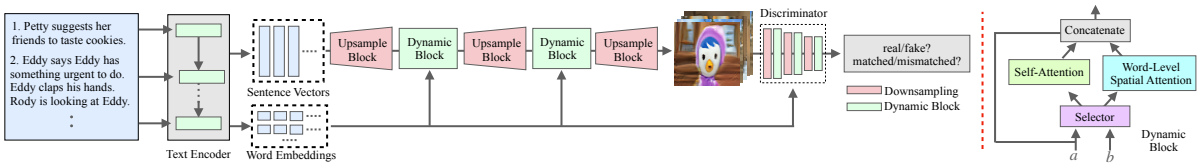

Figure 1: Architecture of the proposed method (left) and dynamic block (right).

are shown in Section 3.1.2). Note that both types of attention share the same attention weights, and SA mainly focuses on capturing correlations between long-range pixels within the same image (Li and Lukasiewicz, 2022a), and WSA mainly focuses on fusing cross-domain text information with intermediate image features in both the generator and the discriminator at a matched semantic level.

3.1.2 Dynamic Block

Our proposed dynamic block is in Fig. 1, right. Unlike current methods that only allow the interaction between different modal representations at specific locations with fixed semantics, we propose the dynamic block to learn to achieve a dynamic interaction in an end-to-end manner. Our dynamic block is based on a attention selector, which selectively chooses an appropriate attention (i.e., SA or WSA) to further explore semantic depth or fuse these cross-domain pieces of information together.

To achieve the selection effect, we first get global representations $\bar{a}$ and $\bar{b}$ of both a and b (see Section 3.1.1) using average pooling. Then, the correlation w between a and b can be obtained via $w = \text{Sigmoid}(\bar{a}\bar{b})$, where w denotes the correlation level between a and b. Then, Gumble-Softmax reparameterization (Jang et al., 2016) is adopted to choose a particular attention, based on the probability of each attention, e.g., the probability of SA can be defined as:

$$p(\text{SA}) = \frac{\exp((\log(w) + z)/\tau)}{\exp((\log(w) + z)/\tau + \exp((\log(1-w) + z)/\tau}, \quad (1)$$

where $z = -\log(-\log(\mu))$ is sampled Gumble noise, μ is drawn from the uniform distribution, and τ is a hyperparameter. Similarly, $p(\text{WSA})$ is obtained from $p(\text{SA})$ by replacing w with $1-w$ in the numerator. So, given a and b, our dynamic block performs soft weighting in training and hard selection at inference, denoted as:

$$h_{\text{soft}} = p(\text{SA})h_{\text{SA}} + p(\text{WSA})h_{\text{WSA}} \quad (2)$$

$$h_{\text{hard}} = \begin{cases} h_{\text{SA}}, & \text{if } p(\text{SA}) > p(\text{WSA}) \\ h_{\text{WSA}}, & \text{if } p(\text{SA}) \leqslant p(\text{WSA}), \end{cases} \quad (3)$$

where h_{SA} denotes using self-attention to further explore semantic depth, and h_{WSA} denotes fusing finer word information into the generation pipeline.

3.2 Objective Functions

The training follows the training procedure of GANs, where the generator and discriminator are trained alternatively by minimizing their losses.

4 Experiments

To evaluate the performance of our approach, we compare it with StoryGAN (Li et al., 2019c), CP-CSV (Song et al., 2020), DUCO (Maharana et al., 2021), and VLC (Maharana and Bansal, 2021).

4.1 Implementation

We evaluated our approach at the resolution 64×64. The text encoder is a bi-directional LSTM, pre-trained to maximize the cosine similarity between matched image and text features (Xu et al., 2018). We selected the best checkpoints and tune hyperparameters by using the FID and FSD scores. The network was trained for 120 epochs on both Pororo-SV and Abstract Scenes. The Adam optimizer (Kingma and Ba, 2014) was adopted with learning rate 0.0002. We evaluated our approach on a single Quadro RTX 6000 GPU.

4.2 Datasets

Pororo-SV is adopted to evaluate our approach, which is built on PororoQA, a dataset for video question answering (Kim et al., 2017). In Pororo-SV, each story has five consecutive images with corresponding text descriptions. There are $13,000$ story samples in the training set, and $2,336$ story samples in the test set. Differently, we do not evaluate our approach on CLEVR-SV (Li et al., 2019c), as there are only 15 different words in the entire CLEVR-SV dataset, which might fail to fully explore the multimodal network for the story visualization task. We adopt Abstract Scenes (Zitnick and Parikh, 2013) to further evaluate our approach. Abstract Scenes was proposed for studying semantic information, which contains over $1,000$ sets of 10 semantically similar scenes of children playing outside. The scenes are composed of 58 clip-art objects, and there are six sentences describing different aspects of a scene. In this dataset, we treat scenes from the same set as a story, as they are all

Table 1: Quantitative comparison between different methods on Pororo-SV and Abstract Scenes. For FID, FSD, and the number of parameters, lower is better. For Cosine, higher is better.

	Pororo-SV			Abstract Scenes			Number of Trainable Parameters	
Method	FID↓	FSD↓	Cosine↑	FID↓	FSD↓	Cosine↑	Generator↓	Discriminator↓
StoryGAN	78.64	94.53	0.22	135.16	55.80	3.59	47.0M	47.2M
CP-CSV	67.76	71.51	0.32	-	-	-	86.9M	70.9M
DUCO	95.17	171.70	0.08	142.34	49.16	3.95	53.2M	47.2M
VLC	94.30	122.07	0.21	-	-	-	54.5M	47.2M
Ours	**52.87**	**55.20**	**4.61**	**74.12**	**20.07**	**7.28**	**25.1M**	**23.5M**

created from the same seed scene, sharing similar semantic information.

4.3 Evaluation Metrics

The Fréchet inception distance (FID) (Heusel et al., 2017) and the Fréchet story distance (FSD) (Song et al., 2020) are adopted as quantitative evaluation metrics to evaluate the performance of our approach. FID computes the Fréchet distance between the distribution of real images and the distribution of fake images. Differently from FID focusing on single image, FSD is proposed for the story visualization task, which takes the sequence of images into account. FSD is built on the principle of FID by using $\mathrm{R}(2+1)$ (Tran et al., 2018) as backbone model, where $\mathrm{R}(2+1)$ has a flexible sequence length and the strong ability to capture temporal consistency.

However, as both FID and FSD cannot reflect the semantic alignment between sentences and story images, following (Li et al., 2022a; Li and Lukasiewicz, 2022b), we compute the average cosine similarity (Cosine) between pairs of sentence and synthetic image over the testing set, and further scale the value by 100.

Besides, we show the number of parameters in the generator and the discriminator for different methods on Pororo-SV to compare the size of different networks.

4.4 Qualitative Evaluation

Figs. 2 and 3 show examples of a visual comparison between our approach and the baselines on Pororo-SV and Abstract Scenes, respectively. Our approach generates realistic images with better regional details, text-image alignment, and consistency, for example, shown in Fig. 2, the characters Pororo (i.e., penguin) and Crong (i.e., frog) have a sharper shape with fine-grained regional details, such as hats and glasses, and shown in Fig. 3, our

Figure 2: Qualitative comparison on Pororo-SV.

approach generates a ball, aligned with the given first sentence, while other methods fail to generate the required ball object on the grass.

4.5 Quantitative Evaluation

Table 1 shows a quantitative comparison with the following widely used metrics on Pororo-SV and Abstract: FID (Heusel et al., 2017), FSD (Song et al., 2020), and Cosine (Li et al., 2022a). From the tables, we can observe that our approach achieves better results against others. This illustrates that our approach can generate images with finer quality, achieve better image-text semantic alignment, and keep higher consistency across story images.

We further compare the number of trainable parameters in different methods. As our method is a single-GAN based network, compared to StoryGAN, it reduces the size of the generator by about 46.59%, and of the discriminator by about 50.21%.

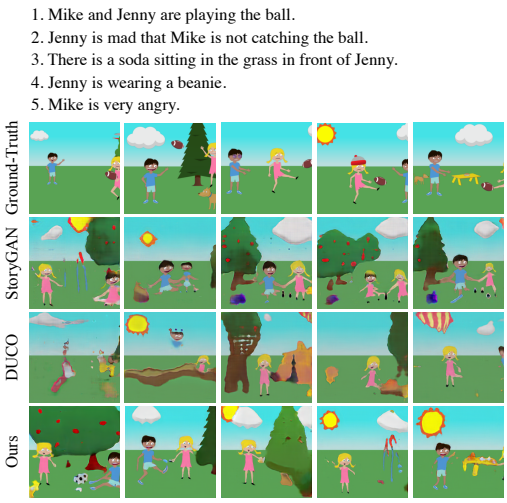

Figure 3: Qualitative comparison on Abstract Scenes.

4.6 Component Analysis

Table 2 shows a component analysis to evaluate the effectiveness of different components. Without either attention in the dynamic block, the performance of our model degrades, and the worst performance is obtained when the model is without using the entire dynamic block. This demonstrates that (1) each attention mode is important in the dynamic block, and (2) simply fusing different-modal representations without considering their semantics fails to comprehensively improve the performance.

Besides, we further consider the effectiveness of the dynamic block in the generator and discriminator. The worse performance in "Ours w/o DB in D" shows that the dynamic block plays a more important role in the discriminator. We think this is because the discriminator needs to provide training feedback to the generator, in terms of image quality and text-image alignment. If the discriminator fails to match the semantics between text and image representations, training feedback may be less precise, which may hinder the generator to generate high-quality story images. The degraded performance in "Ours w/o DB in G" is because the generator cannot effectively capture the correlation between text and image representations, even though there is precise and fine-grained training feedback provided by the discriminator. This demonstrates the complementary effect between the adoption of the dynamic block in both the generator and discriminator.

4.7 Human Evaluation

Similarly to (Li et al., 2022a), a human evaluation on Pororo-SV is conducted based on three evaluation criteria: (1) visual quality, (2) text-image semantic alignment, and (3) consistency across story images. We asked workers to decide which sample is the best, where each sample contains story images and corresponding sentences. 100 randomly selected samples are assigned to three workers to reduce the human variance. Workers prefer the results that are generated by our approach.

Table 2: Ablation study on Pororo-SV. "Ours w/o both Attn" denotes without using both attention in the dynamic block (DB); "Ours w/ SA Only" denotes only adopting the self-attention in DB; "Ours w/ WSA Only" denotes only adopting the word-level spatial attention in DB; and "w/o DB in G (or D)" denotes without using DB in the generator (or discriminator).

Method	FID	FSD
Ours w/o both Attn	69.84	74.25
Ours w/ SA Only	63.87	70.96
Ours w/ WSA Only	60.96	65.82
Ours w/o DB in G	57.10	58.93
Ours w/o DB in D	61.81	60.49
Ours	52.87	55.20

Table 3: Human evaluation on Pororo-SV between VLC, DUCO, and Ours based on three criteria.

Choice (%)	Ours	VLC	DUCO
Visual Quality	80.33	12.00	7.67
Alignment	77.33	13.67	9.00
Consistency	82.00	9.67	8.33

5 Conclusion

In this paper, we explored the semantic misalignment problem existing in current story visualization methods, and further proposed dynamic interactions via learning to dynamically explore various semantic depths and fuse the different-modal information at a matched semantic level. Experiments demonstrate the superior performance of our proposed single-GAN based approach, with a fewer number of parameters, neither using segmentation masks nor auxiliary captioning networks.

Acknowledgments

This work was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, by the AXA Research Fund, and by the EPSRC grant EP/R013667/1. We also acknowledge the use of the EPSRC-funded Tier 2 facility JADE (EP/P020275/1) and GPU computing support by Scan Computers International Ltd.

Limitations

Our method may fail to generate high-quality results when a given multi-sentence story is complex, e.g., it describes multiple characters (e.g., >3) with various backgrounds. Currently, similarly to current methods, our approach focuses more on small-size story image generation, which means that when the number of images in a story is larger (e.g., >15), our method may fail to ensure consistency between different story images.

Ethical Considerations

The datasets that we use in this paper do not have any personally identifiable information or offensive content, as they are cartoon datasets for educational purposes (Pororo-SV and Abstract Scenes).

References

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637.

Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2019. Semantic object accuracy for generative text-to-image synthesis. *arXiv preprint arXiv:1910.13321*.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. DeepStory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Bowen Li and Thomas Lukasiewicz. 2022a. Lightweight long-range generative adversarial networks. *arXiv preprint arXiv:2209.03793*.

Bowen Li and Thomas Lukasiewicz. 2022b. Word-level fine-grained story visualization. *arXiv preprint arXiv:2208.02341*.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019a. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, pages 2063–2073.

Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. 2020. Image-to-image translation with text guidance. *arXiv preprint arXiv:2002.05235*.

Bowen Li, Philip HS Torr, and Thomas Lukasiewicz. 2022a. Clustering generative adversarial networks for story visualization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 769–778.

Bowen Li, Philip HS Torr, and Thomas Lukasiewicz. 2022b. Memory-driven text-to-image generation. *arXiv preprint arXiv:2208.07022*.

Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019b. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12174–12182.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019c. StoryGAN: A sequential conditional GAN for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338.

Adyasha Maharana and Mohit Bansal. 2021. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*.

Louis Mahon, Eleonora Giunchiglia, Bowen Li, and Thomas Lukasiewicz. 2020. Knowledge graph extraction from videos. In *19th IEEE International Conference on Machine Learning and Applications*, pages 25–32.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning text-to-image generation by redescription. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1505–1514.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. 2020. Character-preserving coherent story visualization. In *European Conference on Computer Vision*, pages 18–33. Springer.

Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. 2020. DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–842.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2018. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5802–5810.

C. Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016.