Generating Intermediate Representations for Compositional Text-To-Image Generation

Ran Galun
The Hebrew University of Jerusalem
ran.galun@mail.huji.ac.il

Sagie Benaim
The Hebrew University of Jerusalem
sagie.benaim@mail.huji.ac.il

Abstract

Text-to-image diffusion models have demonstrated an impressive ability to produce high-quality outputs. However, they often struggle to accurately follow fine-grained spatial information in an input text. To this end, we propose a compositional approach for text-to-image generation based on two stages. In the first stage, we design a diffusion-based generative model to produce one or more aligned intermediate representations (such as depth or segmentation maps) conditioned on text. In the second stage, we map these representations, together with the text, to the final output image using a separate diffusion-based generative model. Our findings indicate that such compositional approach can improve image generation, resulting in a notable improvement in FID score and a comparable CLIP score, when compared to the standard non-compositional baseline. Our code is available at https://github.com/RANG1991/Public-Intermediate-Semantics-For-Generation

1 Introduction

Recently, text-to-image generation has shown highly impressive results, primarily using diffusion modeling. To enable effective conditioning, one often integrates textual embeddings as input to the denoising network [1–4]. However, text prompts may fail to enable full and precise spatial control. Describing semantic properties, such as the segmentation of individual objects, or physical characteristics, such as the depth of objects within an image, using solely textual descriptions, can be challenging and inefficient. Consequently, text-to-image diffusion models often need to implicitly infer these properties from the text-image data used during training. This process is prone to inaccuracies and may lead to difficulties in accurately capturing intermediate representations, such as object segmentation or depth. In this study, therefore, we ask whether a compositional approach of first explicitly generating these intermediate representations and subsequently using them as an additional condition can help mitigate these issues and improve text-to-image generation.

Some recent work considered the ability to condition the generation process on some intermediate representation in addition to text [5–8]. However, these approaches require having existing intermediate representations (such as a segmentation map or a depth map) as input, which is not the case in text-to-image generation. Another approach is to first generate an intermediate layout or blobs representation from an input text, and then generate the entire image using the layout or blobs representation and the input text. This can be done by using cross-attention [9–11], or a separate diffusion model [12]. However, since both a layout and a blob cannot encompass fine details, these approaches provide only partial control over the output image.

Instead, we propose to employ a two-step compositional process: (1). First, we generate fine-grained intermediate representations (such as a depth map or a segmentation map) conditioned on the input text. To do so, we fine-tune a pre-trained Stable-Diffusion [4] model to generate an intermediate representation given a text prompt. In the case where more than one representation is generated,

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

we propose an approach to align those representations, making sure that they correspond to the same output image. (2). We then generate the output image conditioned on the input text and the intermediate representations generated. We do so using a pre-trained ControlNet [5], which was trained to generate an output image given both the text and the generated intermediate representations.

We consider three distinct intermediate representations for our experiments: a depth map, a segmentation map, and Hough lines (HED). We assess their impact on text-to-image generation. Our findings indicate that, among these single representations, using either the depth map or the segmentation map (solely, without alignment) as intermediate representations results in a notable improvement in Fréchet Inception Distance [13] (FID) score compared to standard non-compositional Stable-Diffusion baseline. In addition, we explore the generation capabilities using two aligned intermediate representations, revealing insights into their effectiveness and potential benefits.

2 Related Work

Conditional Text-to-Image Generation Early advances were dominated by Generative Adversarial Networks (GANs) [14–16] and later approaches considered an autoregressive approach [1, 17, 18]. Recently, diffusion models [19, 20, 3, 21] have achieved significant improvements. To enable conditional generation, Stable-Diffusion [4] incorporates conditions (e.g. a text prompt) by first encoding them and then applying cross-attention with the denoiser's layers. At inference time, classifier guidance [20] can be used to guide the noise trajectory using an external classifier. Alternatively, classifier free guidance [22], combines the output of an conditional and unconditional model. ControlNet [5] introduces replicated U-Net layers that share weights with the original Stable-Diffusion backbone U-Net. These replicated U-Net layers get as input a control image (e.g. a segmentation map).
Compositional Text-to-Image Generation Several studies considered text-to-image generation as a compositional approach, where first, the condition is generated, and only then the final image is generated based on this condition. Approaches used mainly Large-Language-Models (LLMs) to generate an intermediate representation (e.g., layout, blob) [9, 12, 11], but these do not enable fine-grained control. [23] used a bounding box representation instead. [24] extracts readouts from intermediate features and guides the generation process based on a user input and the readouts. However, these methods still require user input and otherwise cannot achieve fine-grained control.

3 Method

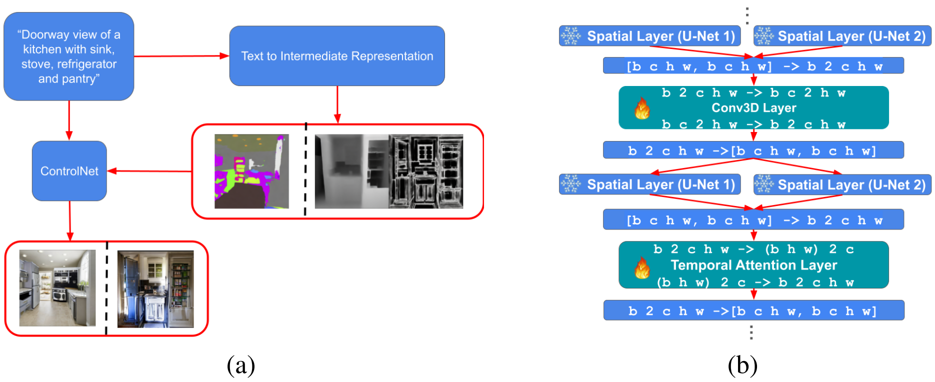
We describe here our two-step compositional generation approach, as illustrated in Fig. 1.

3.1 Generating Intermediate Representations

To generate an intermediate representation given an input text, we consider a text-representation pairs dataset and fine-tune a text-to-image pre-trained Stable-Diffusion (SD) model on this dataset. As the SD's VAE was trained on images, we also fine-tune it on the intermediate representation.

Aligned Intermediate Representations Our fine-tuned SD models can now be used to generate multiple intermediate representations. There is, however, a problem, since these different intermediate representations may not be aligned. While we describe the alignment procedure of two intermediate representations here, this can be extended to a variable number of such representations. Our alignment procedure is inspired by that of [25], which tackles a different problem of aligning the latents of a text-to-image diffusion model to enable text-to-video generation. In particular, we consider two unaligned pre-trained models, one for each intermediate representation. We assume Stable-Diffusion v2.1 model and denote the spatial layers within the U-Net (in both the encoding and decoding paths) of each model as $l^i_{\Theta_1}$ and $l^i_{\Theta_2}$ respectively, where i refers to the layer index.

We now introduce a joint temporal layer, l^i_{Φ}, between consecutive spatial layers. We assume $z^i_{crl_r} \in \mathbb{R}^{B\times C\times H\times W}$ is the output of $l^i_{\Theta_r}$ $(r = 1, 2)$, where C represents the number of latent channels, and H and W denote the spatial latent dimensions. We then concatenate $z^i_{crl_1}$ and $z^i_{crl_2}$ along a new dimension t, referred to as the "temporal" dimension ($t = 2$), resulting in z^i_{crl}. z^i_{crl} passes through two types of temporal mixing blocks: (i) temporal attention layers and (ii) residual blocks employing 3D convolutions. For (i), we used a block denoted as $f_{cross-attn}(z^i_{crl}, c)$, that is defined as follows:



(a) (b)

Figure 1: (a). **Illustration of the full pipeline**. In the first step, we generate aligned intermediate representation(s) given the input text. In the second stage, we use a pre-trained ControlNet to map the input text and the generated intermediate representation(s) to an output image. (b). **Illustration of our alignment procedure**. Given two pre-trained text-to-intermediate models (e.g., text-to-depth and text-to-segmentation), we interleave their spatial layers using "temporal" layers. The "temporal" layers consist of either a 3D convolution or a temporal attention layer and indicate the dimension on which the attention or convolution is performed. For clarity, we also provide each component's input and output dimensions. We note that only the temporal layers are trained in this stage.

$$f_{cross-attn-1} = \text{cross-attn}(\text{lin}_1(z^i_{crl}), \text{lin}_2(c))$$
$$f_{l-norm-1} = \text{l-norm}(\text{lin}_1(z^i_{crl}) + f_{cross-attn-1})$$
$$f_{cross-attn-2} = \text{cross-attn}(f_{l-norm-1}, \text{lin}_2(c))$$
$$f_{l-norm-2} = \text{l-norm}(f_{l-norm-1} + f_{cross-attn-2})$$
$$f_{cross-attn}(z^i_{crl}, c) = \text{lin}_3(f_{l-norm-2})$$

where $c \in \mathbb{R}^{1\times 1024}$ is the text CLIP embedding. lin_i's are linear projection layers and l-norm is a layer norm. Cross-attention is applied as in [26] between a batch of $(B \cdot H \cdot W)$ vectors with a sequence length of 2 or 1. Specifically, the queries in the cross-attention computation are the projected spatial outputs of the two U-Nets, and the keys and values are the projected text embeddings.

Method	↓ FID	↑ CLIP
SD v2.1 (Baseline)	23.44	**30.58**
Ours (Seg)	**19.92**	30.43
Ours (Depth)	20.73	30.30
Ours (HED)	27.56	29.70
Ours (Depth & HED)	50.56	28.88
Ours (Depth & Seg)	32.53	29.82
ControlNet - GT Seg	16.80	30.42
ControlNet - GT Depth	16.17	30.35

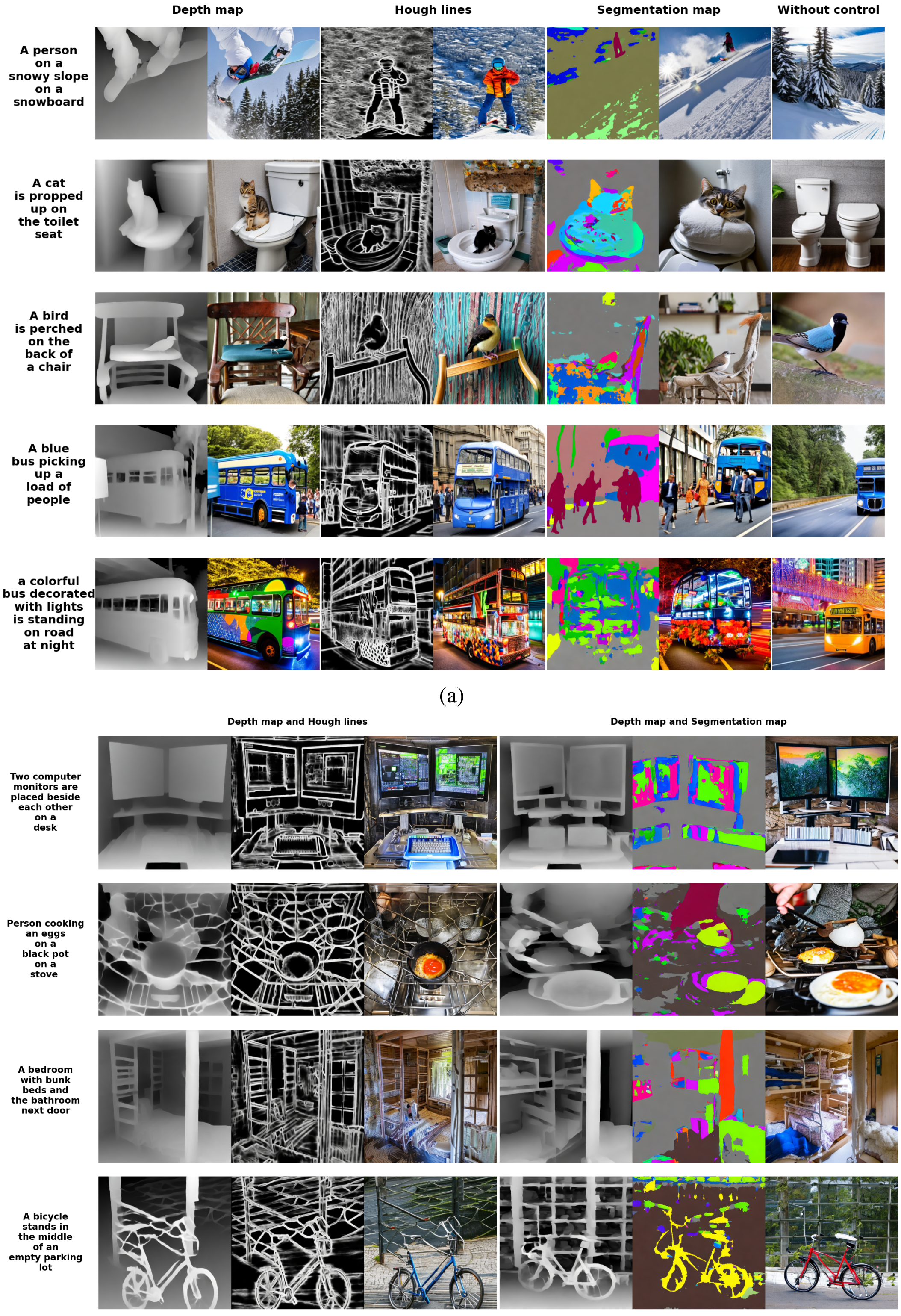
Table 1: FID and CLIP alignment scores. In brackets we note the type of the intermediate representation(s). The bottom two rows provide a comparison when the second stage is used with ground truth intermediate representation. While this is not a direct comparison (as it uses additional input), it provides an upper bound.

For (ii), we used the following block: $f_{conv}(z^i_{crl}) = \text{ReLU}(z^i_{crl} + \text{Conv3D}(z^i_{crl}))$. The input to the 3D convolution block is of shape $B \times C \times 2 \times H \times W$, and the output is of shape $B \times C \times 2 \times H \times W$. That is, we apply 3D convolution on B 4-dimensional tensors. For the convolution parameters, we used a kernel size of $(3, 1, 1)$ and a stride of 1.

Following either temporal mixing blocks (i) or (ii), we apply a residual operation: $\alpha_i \cdot z^i_{crl} + (1 - \alpha_i) \cdot f$, where $\alpha_i = sigmoid(x)$ is a learned value between 0 and 1, and f is either $f_{att}(z^i_{crl}, c)$ or $f_{conv}(z^i_{crl})$. The temporal blocks are trained using standard SD reconstruction objectives on the output representations, given an input text.

3.2 Generating the final output image

Given the previous stage, we can now generate a set of aligned intermediate representations conditioned on the input text. The second stage involves training a ControlNet [5] on a dataset of representation(s)-image pairs. To obtain such pairs we apply off the shelf method for obtaining such representations from images (such as depth estimation or segmentation). We then follow ControlNet's procedure of first training a ControlNet model for each intermediate representation in isolation and then combining them together to enable conditioning by all intermediate representations. The underlying SD model is still conditioned on text in the standard manner as in SD. Once these models are trained, we simply generate the output image by using our generated intermediate representations and input text as a condition to the pre-trained ControlNet model.


Depth map
Hough lines
Segmentation map
Without control
A person on a snowy slope on a snowboard
A cat is propped up on the toilet seat
A bird is perched on the back of a chair
A blue bus picking up a load of people
a colorful bus decorated with lights is standing on road at night
Depth map and Hough lines
Depth map and Segmentation map
Two computer monitors are placed beside each other on a desk
Person cooking an eggs on a black pot on a stove
A bedroom with bunk beds and the bathroom next door
A bicycle stands in the middle of an empty parking lot

(a)

(b)

Figure 2: (a). Results using a single intermediate representation (first six columns) and from original SD (last column). The generated intermediate representation is to the left of each output image. (b). As in (a), but using our aligned intermediate representations (depth & HED or depth & segmentation).

4 Experiments

Implementation Details and Datasets For the training phase, we fine-tuned SD v2.1 models on the intermediate representations (i.e. Depth map, Segmentation map, Hough lines) extracted from the first $300,000$ samples from MS-COCO [27] 2017 training set. We used Uniformer [28] for segmentation estimation, ControlNet [5] implementation for HED generation and Depth Anything [29] for depth estimation. We trained each of our models for 12 epochs on 40GB GPUs, with a learning rate of $1e-5$ and a batch size of 32. We used AdamW optimizer with a weight decay of 0.01. For the evaluation phase, we followed previous papers and evaluated our models on $25,000$ samples from MS-COCO 2017 validation set. For the sampling process, we used 80 DDIM steps.

Using A Single Intermediate Representations Tab.1 provides a numerical evaluation of FID and CLIP similarity scores in comparison to SD baseline. Our compositional approach achieves lower FID scores than the original non-compositional SD model, for both the depth and segmentation intermediate representations, except for HED. We hypothesize that this is a result of the domain shift that occurs between the generated Hough lines in contrast to the real Hough lines. As the second stage trains on such realistic data, this domain shift can result in significant errors. Fig. 2(a), provides a visual illustration of outputs of our model in comparison to the SD baseline, showing examples whereby predicting intermediate representation improves overall text-image correspondence.

Using Aligned Intermediate Representations Tab. 1 and Fig. 2(b) present the corresponding numerical and visual results when using aligned intermediate representations. While our method produces aligned outputs, we observe a drop in performance, both in terms of FID and CLIP scores. We hypothesize that, while our alignment model produces aligned outputs, this comes at the expense of quality. This quality degradation results in inputs which are far from real intermediate representation, resulting in a domain shift, which ultimately results in worse performance in the second stage when these intermediate representations are fed into ControlNet.

5 Conclusion

In this work, we proposed a two-stage compositional approach for text-to-image generation, comprising of first generating intermediate representations and subsequently using these representations to generate a final output image. Our compositional apporach demonstrated improved FID scores over the non-compositional baseline when using a single depth or segmentation maps as intermeidate representations. Future work could focus on refining the alignment process, and addressing the domain shift that occurs between generated representations and those used as input in the second stage of our compositional approach.

References

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

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

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

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

[5] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[6] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:4296–4304, Mar. 2024. doi: 10.1609/aaai.v38i5.28226.

[7] Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, and Jingfeng Zhang. Scedit: Efficient and controllable image diffusion generation via skip connection editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8995–9004, 2024.

[8] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023.

[9] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

[10] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7942, 2024.

[11] Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. *arXiv preprint arXiv:2405.08246*, 2024.

[12] Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Dynamic equilibrium between realism and compositionality improves text-to-image diffusion models. *arXiv preprint arXiv:2402.12908*, 2024.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

[15] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[16] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 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, 2018.

[17] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

[18] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[23] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5544–5552, 2024.

[24] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8217–8227, 2024.

[25] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[26] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600, 2023.

[29] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.