Vision + Language Applications: A Survey

Yutong Zhou, Nobutaka Shimada
Ritsumeikan University, Shiga, Japan
zhou@i.ci.ritsumei.ac.jp

Abstract

Text-to-image generation has attracted significant interest from researchers and practitioners in recent years due to its widespread and diverse applications across various industries. Despite the progress made in the domain of vision and language research, the existing literature remains relatively limited, particularly with regard to advancements and applications in this field. This paper explores a relevant research track within multimodal applications, including text, vision, audio, and others. In addition to the studies discussed in this paper, we are also committed to continually updating the latest relevant papers, datasets, application projects and corresponding information at https://github.com/Yutong-Zhou-cv/Awesome-Text-to-Image.

1. Introduction

> *"The baby, assailed by eyes, ears, nose, skin, and entrails at once, feels it all as one great blooming, buzzing confusion."*
>
> – William James (1890)

The human perceptual system is a complex and multifaceted construct. The five basic senses of hearing, touch, taste, smell, and vision serve as primary channels of perception, allowing us to perceive and interpret most of the external stimuli encountered in this "blooming, buzzing confusion" world. These stimuli always come from multiple events spread out spatially and temporally distributed. The human brain processes conceptual representation based on various modes of perception, with visual information accounting for as much as 87%. Furthermore, in cases where certain aspects of content cannot be only conveyed through visual information, combining multiple sensory modalities can potentially provide a more comprehensive understanding of concepts. In other words, we constantly perceive the world in a "multimodal" manner, which combines different information channels to distinguish features within confusion, seamlessly integrates various sensations from multiple modalities and obtains knowledge through our experiences.

Figure 1. **Overall structure of our survey.** Zoom in for details.

In the last few decades, Computer Vision (CV) and Natural Language Processing (NLP) have made several major technological breakthroughs in deep learning research. In recent years, in addition to the significant advancements made in single-modality models [16, 181, 206], there has been an upsurge in research activities focused on developing large-scale multimodal approaches [41, 125, 146]. Despite the impressive advancements in text-to-image synthesis and its related tasks, which effectively incorporate multiple modalities, existing literature surveys [1, 13, 19, 45, 54, 130, 207, 208, 222] still remain lacking to thoroughly review and summarize the progress of the applications and challenges in this field. As illustrated in Fig. 1, this review complements previous surveys on the text-to-image synthesis task with a focus on representative algorithms and expanded applications beyond text-to-image. Our survey aims to accomplish several goals: (1) Present a comprehensive reference that covers the current state-of-the-art research progress and practical breakthroughs in multimodal learning centering on vision and language, (2) Provide a valuable resource for novice researchers seeking to explore this field, (3) Investigate compounding issues and business analysis in Artificial Intelligence-Generated Content (AIGC) field, and (4) Inspire future research directions and upcoming advances through emerging trends and challenges.

Table 1. **Chronological timeline of representative text-to-image datasets.** *"Public"* includes a link to each dataset (if available) or paper (if not). *"Annotations"* denotes the number of text descriptions per image. *"Attrs"* denotes the total number of attributes in each dataset.

Year	Dataset	Public	Details				
			Category	Images (Resolution)	Annotations	Attrs	Other Information
2008	Oxford-102 Flowers [124]	✓	Flower	8,189 (-)	10	-	-
2011	CUB-200-2011 [179]	✓	Bird	11,788 (-)	10	-	BBox, Segmentation...
2014	MS-COCO2014 [109]	✓	Iconic Objects	120k (-)	5	-	BBox, Segmentation...
2018	Face2Text [50]	✓	Face	10,177 (-)	1~	-	-
2019	SCU-Text2face [24]	✗	Face	1,000 (256×256)	5	-	-
2020	Multi-Modal CelebA-HQ [196]	✓	Face	30,000 (512×512)	10	38	Masks, Sketches
2021	FFHQ-Text [223]	✓	Face	760 (1024×1024)	9	162	BBox
2021	M2C-Fashion [217]	✗	Clothing	10,855,753 (256×256)	1	-	-
2021	CelebA-Dialog [77]	✓	Face	202,599 (178×218)	~5	5	Identity label...
2021	Faces a la Carte [186]	✗	Face	202,599 (178×218)	~10	40	-
2021	LAION-400M [159]	✓	Random Crawled	400M (-)	1	-	KNN index...
2022	Bento800 [225]	✓	Food	800 (600×600)	9	-	BBox, Segmentation, Label...
2022	LAION-5B [158]	✓	Random Crawled	5.85B (-)	1	-	URL, Similarity, Language...
2022	DiffusionDB [189]	✓	Synthetic Images	14M (-)	1	-	Size, Random seed...
2022	COYO-700M [18]	✓	Random Crawled	747M (-)	1	-	URL, Aesthetic Score...
2022	DeepFashion-MultiModal [78]	✓	Full body	44,096 (750×1101)	1	-	Densepose, Keypoints...
2023	ANNA [144]	✓	News	29,625 (256×256)	1	-	-
2023	DreamBooth [153]	✓	Objects & Pets	158 (-)	25	-	-

2. Background

This section introduces commonly used datasets and evaluation metrics of the text-to-image generation task.

2.1. Datasets

As one might expect, NLP models primarily rely on textual data for training, while CV models train on image-based information. Vision-language pre-trained models employ a combination of text and images, merging the capabilities of both NLP and CV. Training complex models requires accurate annotation of data by expert human annotators. This process incurs considerable costs and resources, particularly when working with large datasets or domain-specific information. Table 1 presents a comprehensive list of notable datasets, ranging from small-scale [24, 50, 124, 144, 153, 179, 223, 225] to large-scale [18, 77, 78, 109, 158, 159, 186, 189, 196, 217], within the field of text-to-image generation research.

[109, 124, 179] are widely known in the early stages of text-to-image generation research and regarded as the most commonly utilized dataset in the field. [223] is a small-scale collection of face images with extensive facial attributes specifically designed for text-to-face generation and text-guided facial image manipulation. [225] is the first manually annotated synthetic box lunch dataset containing diverse annotations to facilitate innovative aesthetic box lunch presentation design. [78] is a high-quality human dataset annotated with rich multi-modal labels, including human parsing labels, keypoints, fine-grained attributes and textual descriptions. As the pioneer in large-scale text-to-image datasets, [189] comprises 14 million images generated through Stable Diffusion [150] using prompts and hyperparameters provided by real users. Furthermore, [158] contains 5.85 billion CLIP-filtered image-text pairs, the largest publicly available image-text dataset globally.

2.2. Evaluation Metrics

When evaluating the performance of text-to-image generation models, two assessments are commonly employed: automated evaluation and human evaluation. The former measures image realism and the alignment between generated images and corresponding text descriptions. The latter relies on humans to make subjective judgments.

Automatic Evaluation Automatic evaluation is a widely used approach for assessing the quality of generated images and quantitatively measuring the alignment between the generated images and input textual descriptions.

For image quality assessment, Inception Score (IS) [156] and Fréchet Inception Distance (FID) [63] are two commonly utilized metrics. IS evaluates the quality and diversity of the generated images. *Higher is better.* FID measures the Fréchet distance between the generated images and the ground truth images. *Lower is better.*

For text-image consistency assessment, R-precision (RP) [199] evaluates the degree of alignment between a generated image and its corresponding text description. *Higher is better.* Semantic object accuracy (SOA) [64] utilizes a pre-trained object detector to check whether the generated image contains objects mentioned in the corresponding textual description. User studies have demonstrated that SOA shows a stronger correlation with human perception than IS. *Higher is better.* Positional Alignment (PA) [34] evaluates the consistency between the spatial placement of the visual elements within the generated image and their corresponding positions in the text description. *Higher is better.*

Although automatic evaluation metrics have demonstrated effectiveness in assessing the quality of generated images in text-to-image synthesis, these metrics are subject to several limitations. One of the most significant issues is that such metrics may attribute high scores to images that appear realistic but fail to represent the intended meaning of the corresponding textual description accurately. Moreover, automatic metrics typically neglect the subjective nature of human perception, which plays a crucial role in evaluating the quality of generated images. Therefore, complementary human evaluation metrics are necessary to provide a comprehensive assessment of the quality of generated images in text-to-image synthesis.

Human Evaluation To confirm the credibility and correlation of automatic evaluations, several works [101, 108, 131, 138, 155] have conducted user analysis to evaluate the performance of generated images against human judgments. In [34, 205], participants are asked to rate generated images based on two criteria: plausibility (including object accuracy, counting, positional alignment, or image-text alignment) and naturalness (whether the image appears natural or realistic). The evaluation protocol is designed in a 5-Point Likert [141] manner, in which human evaluators rate each prompt on a scale of 1 to 5, with 5 representing the best and 1 representing the worst. A human evaluation in [138] was conducted for three challenging tasks to compare the performance of Stable Diffusion [150] and DALL-E 2 [145]. Moreover, for rare object combinations that require common sense understanding or aim to avoid bias related to race or gender, human evaluation is even more important. Such concepts are difficult to define, and human expertise is required to provide accurate evaluations.

3. Generative Models

3.1. GAN-based Model

Generative Adversarial Networks (GANs) [52] have been a significant breakthrough in generative modeling, with the unique ability to generate realistic and diverse samples. GANs consist of two components, the generator and the discriminator, which engage in a continuous adversarial process. The generator produces synthetic images from random noise, while the discriminator aims to distinguish between these generated images and real images from the training dataset. Through backpropagation and optimization, the generator continually refines its outputs in response to the feedback from the discriminator and generates more realistic images. GANs have been applied to many applications, including image generation [9, 81, 86], super-resolution [17, 213], 3D object generation [166, 212], *etc*. From image inpainting [30, 216] to video generation [57, 176], from makeup transfer [76, 200, 201] to virtual try-on [96, 197], GANs solve various problems and create new possibilities across multiple industries.

3.2. Diffusion Model

Diffusion models (DMs), also commonly known as diffusion probabilistic models [167], represent a class of generative models founded on Markov chains and trained through weighted variational inference [66]. The primary objective of [66] is to learn the impact of noise on the available information in the sample or the degree to which the diffusion process reduces the information available. The two-step process in [66] consists of forward and reverse diffusion. In the forward diffusion process, Gaussian noise is successively introduced, representing the diffusion process until the data becomes total noise. The reverse diffusion process trains a neural network to learn the conditional distribution probabilities, allowing the model to reverse the noise and reconstruct the original data effectively. DMs also address some challenges associated with GANs, such as mode collapse and training instability.

4. Generative Applications

In the previous sections, we explore generative models from the perspective of their background and basic mechanisms. This section delves into the applications in multimodal backgrounds, focusing on their utilization in generating and analyzing data. The discussion is organized into subsections, each focusing on a combination of various modalities, particularly in Sec. 4.2 and Sec. 4.3, where "X" represents the additional data. We analyze the significance of integrating multiple modalities for enhanced performance in vision and language tasks. Additionally, we also explore the methodologies utilized in developing these multimodal learning frameworks and discuss their respective contributions to advancing the field.

4.1. Text-to-Image

Advancements in the text-to-image generation domain can be broadly classified into three primary categories: GAN-based methods, diffusion-based methods and autoregressive methods, as illustrated in Tab. 2. GAN-based approaches are primarily based on stackGAN [209] and styleGAN [85] architectures to generate high-quality and visually coherent images. Diffusion-based methods model the image generation process as a series of diffusion steps [167], progressively refining the generated image. Inspired by the success of autoregressive Transformers [177], autoregressive methods focus on sequentially predicting individual pixels or regions in an image, generating the output based on a learned probability distribution. Each approach has unique advantages and challenges in text-to-image synthesis, and further research aims to address their limitations and enhance their capabilities.

Figure 2. **Diverse text-to-face results generated from GAN-based [105, 139, 224] / Diffusion-based [150] / Transformer-based [145] models.** Images in orange boxes are captured from original papers (a) [224], (b) [139] and (c) [105]; others are generated by a pre-trained model [139] [(b) left bottom row], Dreamstudio [(a-c) middle row] and DALL-E 2 [(a-c) right row] online platforms from textual descriptions. Further details about online platforms are mentioned in Sec. 5.2.

Table 2. **A comprehensive list of text-to-image approaches.** The pioneering works in each development stage are highlighted in blue. Text-to-face generation works are underlined.

Models		Years: Methods
GAN (Sec. 3.1)	Conditional GAN [119]-based	**2016-2021:** [144], [148], [20], [35], [67], [132], [38]
	StackGAN [189]-based	**2018:** [210], [53], [199], [160] **2019:** [24], [80], [172], [204], [142], [97], [228] **2020:** [227], [25], [173], [106], [184], [64] **2021:** [170], [223], [224], [202], [36], [26] **2022:** [116], [115], [174], [194]
	StyleGAN [85]-based	**2021:** [139], [186], [183] **2022:** [37], [139], [69], [169], [154], [105], [226]
	Others	**2018:** [218] (Hierarchical adversarial network) **2021:** [44, 113] (BigGAN [14]) **2022:** [107] (One-stage framework)
Diffusion (Sec. 3.2)	Diffusion [167]-based	**2022:** [150], [55], [145], [155], [123], [110] [62], [195], [83], [99], [43], [6], [92] **2023:** [103], [42], [22], [214], [128]
Autoregressive	Transformer [177]-based	**2021:** [144], [32], [73] **2022:** [33], [205], [100], [27], [188], [193], [94], [48]

Text-to-Face Text-to-image technology has been extensively applied in various applications. However, compared to other text-to-image applications, such as text-to-object or text-to-scene synthesis, text-to-face synthesis has received comparatively less attention (see underlined in Tab. 2). As an early text-to-image application, facial visual reconstruction involved artists sketching suspects to help eyewitnesses recall more details. This process depends heavily on accurate descriptions and requires artists to have extensive training in facial sketching techniques. Moreover, text-to-face synthesis has significant potential in human retrieval, public security, and missing child renderings. Consequently, developing text-to-face synthesis techniques requires a meticulous approach to address challenges stemming from ambiguous and complex textual descriptions, thereby enhancing its practical significance. Comparison results for several text-to-face synthesis models are illustrated in Fig. 2.

4.2. Text-to-X

Video Automatically generating videos based on textual descriptions is a highly challenging task, as it requires addressing two critical issues: visual quality and semantic consistency. Visual quality refers to the generation of videos that are both realistic and visually coherent, while semantic consistency refers to the generated video sequence that can accurately represent the textual description.

Early text-to-video generation works based on GANs [5, 31, 87, 102, 104, 129] are limited to generating low-resolution short videos depicting simple scenes. With the ongoing development of transformer and diffusion models, recent methods based on transformers [68, 178] and diffusion [65, 163] improve the video quality (see Fig. 3 (a)) and speed [219]. Uriel *et al*. [164] first propose a method for generating three-dimensional (3D) video from text descriptions. These dynamic scenes can be viewed from any camera location and composited into any 3D environment, breaking barriers between text-to-video and text-to-3D.

3D Recent studies [108, 140, 182] have shown that neural radiance fields (NeRFs) [118] can be effectively optimized using text-to-image diffusion models, leading to the generation of high-quality 3D content from textual descriptions (see Fig. 3 (b)-Left and Fig. 3 (b)-Right). Additionally, Amit *et al*. [143] combine recent developments in personalizing text-to-image models (DreamBooth [152]) with text-to-3D generation (DreamFusion [140]) and present DreamBooth3D, a text-to-3D generative model that can produce high-quality 3D images, which trained from as few as 3-6 casually captured images.

Human Motion In the text-to-X domain, text-to-human motion generation [56, 137, 175, 215] remains a relatively niche field. The generated motions are expected to have diversity, enabling exploration of the text-grounded motion space. More importantly, these motions should be accurately captured and well-adapted to the content from the input textual descriptions (see Fig. 3 (c)).

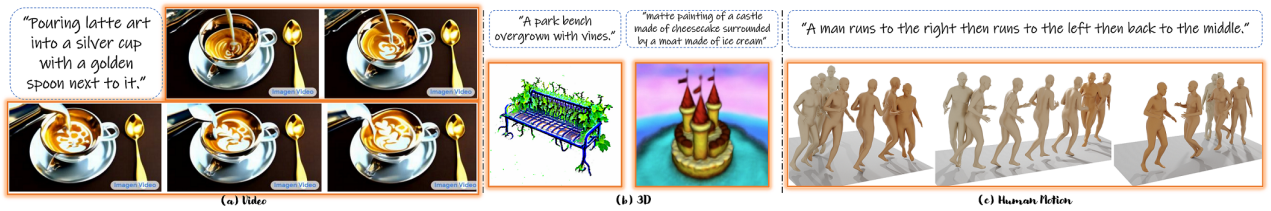

Figure 3. **Selected representative samples on Text-to-X.** Images are captured from original papers ((a) [65], (b)-Left [198], (b)-Right [140], (c) [175]) and remade.

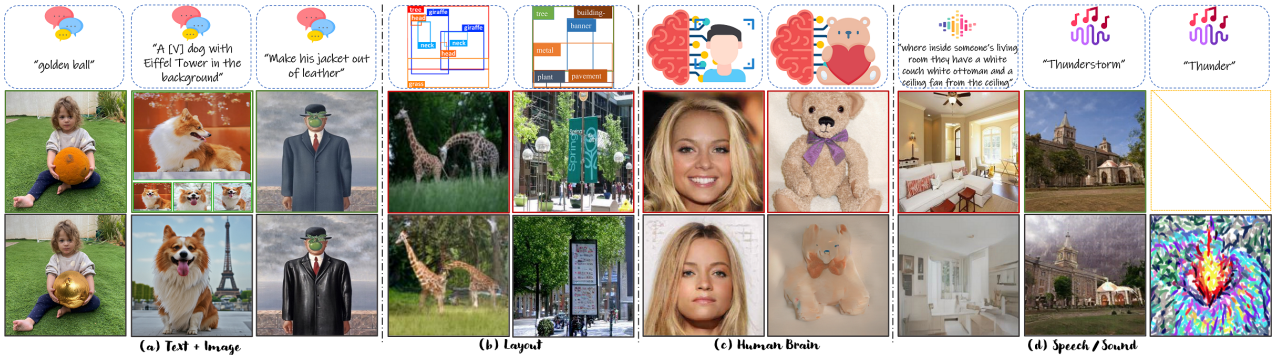

Figure 4. **Selected representative samples on X-to-Image.** Images are captured from original papers and remade. **(a)** Layered Editing [10] (Left), Recontextualization [153] (Middle), Image Editing [15] (Right). **(b)** Context-Aware Generation [61] (Left), Model Complex Scenes [203] (Right). **(c)** Face Reconstruction [29] (Left), High-resolution Image Reconstruction [171] (Right). **(d)** Speech to Image [187] (Left), Sound Guided Image Manipulation [95] (Middle), Robotic Painting [120] (Right). ***Legend:*** *X excluding "Additional Input Image"* (Blue dotted line box, top row). *Additional Input Image* (Green box, middle row). *Ground Truth* (Red box, middle row). *Generated / Edited / Reconstructed Image* (Black box, bottom row).

4.3. X-to-Image

Text + Image Text-to-image generation and text-guided image generation (text+image-to-image) are two distinct techniques within image synthesis. Text-to-image generation models directly create images from a textual description. In this process, the synthesis model is trained to produce images that closely align with the input description, and the generated images are expected to represent the content of the input text visually.

In contrast, text-guided image generation refers to a process where a pre-existing image is modified or manipulated based on a textual input, utilizing textual information to make targeted modifications rather than generating a new image. This approach often involves tasks such as image editing [4, 10, 15, 28, 71, 88, 121, 134, 136, 229] (Samples are shown in Fig. 4 (a)-Left & Right), recontextualization [153] (Results are shown in Fig. 4 (a)-middle), inpainting [122], colorization [51], video generation [39, 39, 47, 70], image stylization [39, 47] or style transfer [46, 79, 93, 157], wherein the textual guidance serves to augment or refine specific elements of the initial image. These advances in text-guided image generation have the potential to revolutionize the way we create and manipulate images, offering new possibilities for digital art, advertising and entertainment.

Layout Layout-to-image generation is a subdomain within image synthesis which focuses on generating coherent and visually consistent images based on a given layout. The layout typically consists of a structured representation of the spatial arrangement of various objects, scenes, or semantic regions within the image [61, 74, 180, 203]. It includes information about the location, size, shape, and relationships of these elements. [61] introduces a context-aware feature transformation module in the generator to ensure the generated encoding accounts for coexisting objects in the scene (see Fig. 4 (b)-Left). [203] compresses RGB images into patch tokens and only focuses on highly-related patch tokens specified by the spatial layout for modeling, thereby achieving disambiguation during the training process (see Fig. 4 (b)-Right).

Human Brain Neural encoding and decoding are two fundamental concepts in neuroscience to make sense of brain-activity data. The former investigates how individual neurons or neural networks represent information through action potentials. The latter involves mapping brain responses to sensory stimuli through feature space. Both fields have a long history of research [75, 90] and are still actively studied with the development of deep learning [11, 161, 162].

Table 3. **Overview of various existing works with multi-modal tasks.**

Year	Method \ Tasks	T2I	T2V	(T+X)2I	LYT2I	SKT2I	SEG2I	I2I	UIG	SR	IC	Other Tasks
2021	UFC-BERT [217]	✔	-	Partial Image	-	-	-	✔	✔	-	-	Multimodal Controls
2021	ERNIE-ViLG [211]	✔	-	-	-	-	-	-	-	-	✔	Generative VQA
2022	OFA [185]	✔	-	-	-	-	-	-	-	-	✔	VQA, ...
2022	Frido [40]	✔	-	-	✔	-	-	-	✔	-	-	SG2I
2022	LDMs [150]	✔	-	-	✔	-	-	-	-	✔	-	Inpainting
2022	NÜWA [191]	✔	✔	Image/Video	-	✔	-	✔	-	-	-	Video Prediction, ...
2022	MMVID [59]	-	✔	Partial Image	-	-	-	-	-	-	-	Multimodal Controls
2022	PoE-GAN [72]	✔	-	SEG/SKT/Image	-	✔	✔	-	-	-	-	(SEG+SKT)2I
2022	AugVAE-SL [89]	✔	-	-	-	-	-	-	-	-	✔	Image Reconstruction
2022	NUWA-Infinity [190]	✔	✔	-	-	-	-	-	✔	-	-	Outpainting(HD), ...
2023	SDG [114]	✔	-	Image	-	-	-	✔	-	-	-	Style-guided, ...
2023	Muse [21]	✔	-	Image	-	-	-	-	-	-	-	Inpainting, Outpainting
2023	MCM [58]	-	-	SEG/SKT	-	✔	✔	-	-	-	-	(SEG+SKT)2I
2023	TextIR [114]	-	-	Image	-	-	-	-	-	✔	-	Inpainting, Colorization
2023	GigaGAN [82]	✔	-	Image	-	-	-	-	-	✔	-	Style Mixing, ...
2023	UniDiffuser [8]	✔	-	Image	-	-	-	-	✔	-	✔	Joint Generation
2023	Visual ChatGPT [192]	✔	-	Image	-	✔	✔	✔	-	-	✔	Edge-to-Image,...

<**Acronym**: Meaning> **T2I**:Text-to-Image; **T2V**:Text-to-Video; **T+X**:Text+X; **LYT**:Layout; **SKT**:Sketch; **SEG**:Segmentation; **UIG**:Unconditional Image Generation; **SR**:Super Resolution; **IC**:Image Captioning/Image-to-Text; **VQA**:Visual Question Answering; **HD**:High Definition; **SG**:Scene Gragh.

Dado *et al*. [29] integrate GANs [52] in the neural decoding of faces. They utilize a pretrained generator of progressive growing GAN (PGGAN) [84] to synthesize photorealistic faces from latents. Meanwhile, a decoding model predicts latents from whole-brain functional Magnetic Resonance Imaging (fMRI) activations (as shown in Fig. 4 (c)-Left). Takagi and Nishimoto [171] demonstrate the ability to decode perceptual content from brain activity by converting it into internal representations of Stable Diffusion [150] without fine-tuning (as shown in Fig. 4 (c)-Right).

Speech / Sound According to the statistics, approximately 50% of the world's languages lack a written form, which makes it impossible to benefit from current text-based technologies. Speech-to-image generation approaches [95,98,120,187] translate speech descriptions into photorealistic images without relying on textual information. The speech embedding network in [187] learns speech embedding with visual supervision, while the generative model synthesizes visually consistent images (see Fig. 4 (d)-Left) based on corresponding speech descriptions.

In addition to the image generation from speech, certain studies also investigate the provision of targeted audio between distinct modalities, such as sound-image [95, 120] (see Fig. 4 (d)-Middle & Right), speech-gesture [2, 91], music-motion [3], music-dance [23, 165], *etc*.

4.4. Multi Tasks

Multimodal learning, incorporating multiple data sources, including text, image, audio, and video, has gained increasing attention due to its ability to solve multiple related tasks simultaneously. Additionally, multi-task learning enables the models to share knowledge across tasks, resulting in improved performance on individual tasks and enhanced generalization capabilities.

In recent years, significant progress has been made in developing multimodal and multi-task learning approaches, as shown in Tab. 3. These architectures facilitate the integration of various modalities by learning shared representations and task-specific features, ultimately leading to more effective and efficient models.

5. Discussion

5.1. Compounding Issues

Computational Aesthetic As AI technology continues to evolve and become more accessible, users are increasingly seeking convenience, better user experience, visual aesthetics, and perceived attractiveness. Computational aesthetic evaluation has been employed for assessing potential AI creativity is much older than text-to-image generation [49]. State-of-the-art text-to-image models increasingly consider aesthetics which attempt to produce better images via better prompts [127].

Since computational aesthetic evaluation remains challenging, human feedback is necessary to determine the optimal prompt formulation and keyword combinations. Pavlichenko and Ustalov [135] adopt the most popular keywords to evaluate their effectiveness in Stable Diffusion [150] and propose a set of keywords that produce images with the highest degree of aesthetic appeal. Jonas [126] emphasizes that the ability to generate effective prompts depends on a thorough understanding of the training set and knowledge of various prompt modifiers. This ability and knowledge together form the field of "prompt engineering".

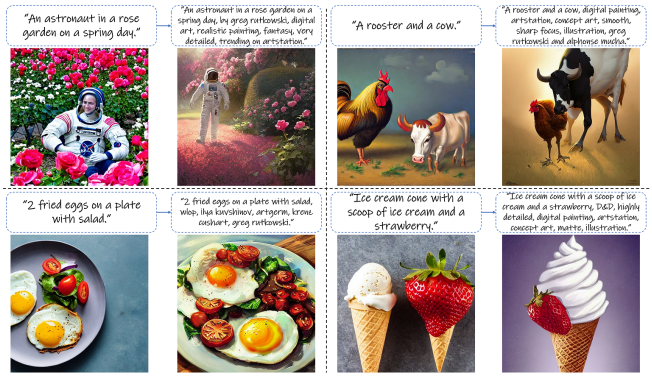

Figure 5. **Sample images generated from user input and optimized prompts.** Images are generated by Stable Diffusion [150], while optimized prompts are generated by Promptist [60].

Prompt Engineering In NLP, prompt-based learning (PBL) [111, 220, 221] employ pre-trained language models on large amounts of text data to address diverse downstream tasks. The process of prompt engineering [149] is crucial for generating task-specific prompt templates, and the effectiveness of PBL greatly depends on the construction method of these templates. The most basic prompt construction method is manual construction, which involves designing appropriate text templates for the target problem, such as translation [16].

In CV, text-to-image generative models present a powerful way to generate realistic images. While using text as input allows for an unlimited range of outputs, users must engage in trial and error with the text prompt when the output is poor quality. Prompt engineering [112] for text-to-image, also referred to as prompt design [149] or prompting, is an emerging technology that utilizes carefully selected sentences to achieve a specific visual style in the synthesized image [151]. Liu and Chilton [112] design five experiments to explore different aspects of prompt engineering for text-to-image generative models, including prompt permutations, random seeds, optimization length, style keywords, and subject and style keywords. They also perform a thorough analysis of the failure and success modes of the above generations. Hao *et al*. [60] propose a prompt adaptation framework (called "Promptist") as an alternative to laborious manual prompt engineering. They use a few manual prompts for supervised fine-tuning and reinforcement learning to generate better prompts. Reinforcement learning involves a reward function that encourages generating more aesthetically appealing images while maintaining the original user intentions. Experimental results (as shown in Fig. 5) indicate that optimized prompts can enhance performance in the following aspects: "Aesthetics augmentation" (top-left), "Content rationalization" (top-right), "Style transformation" (bottom-left), and "Accurate expression" (bottom-right).

Table 4. **Overview of represented online platforms for text-to-image generation.** (Arrange in ascending order of usability and complexity.) DALL-E 2 [145] and Stable Diffusion [150] are the two most commonly utilized models.

Platform	Models	Price	Additional	Links
DeepAI	[150]	Free	Style	Website
Wombo	-	Free	Style	Website
Craiyon	-	Free	-	Website
Yunjing	[150], ...	Free	Chinese prompts	Website
Replicate	[150]	Free	-	Website
Nightcafe	[145], [150], ...	Free	Style	Website
Freehand	-	Free	Chinese prompts	Website
HuggingFace	[150]	Free	-	Website
SD Playground	[150]	Free	-	Website
Bing Image Creator	[145]	Free	-	Website
Lexica	-	Monthly 100 images	Search	Website
starryai	-	Daily 5 free credits	Style, Image+Text	Website
Dreamstudio	[150]	200 free credits	Image+Text	Website
Midjourney	-	20 times free	-	Website
DALL-E 2	[145]	Monthly 15 free credits	-	Website
Firefly	-	Application is required	-	Website

5.2. Business Analysis

Online Platforms As AI-based image generators become a topic of widespread discussion, significant technological advancements have greatly enhanced the accessibility of these tools to the general populace. The emergence of OpenAI's DALL-E [146] and DALL-E 2 [145] marked a crucial milestone in the evolution of AI-based text-to-image generation. While certain text-to-image tools are available for free, others may require a subscription or offer a trial period. Moreover, several platforms provide additional features to enhance the user experience. Table 4 provides an overview of represented online platforms and websites that easily create images from text prompts.

Ethical Considerations Despite the significant progress made in open-source text-to-image generation models, the technology has not yet reached commercial viability due to concerns about unconsciously producing offensive or potentially dangerous biased images. These biases include ambiguity, immorality, stereotypes, or other negative connotations. This is frequently attributed to a lack of consideration for ethical considerations in conventional approaches.

For ***ambiguity issues***, Mehrabi *et al*. [117] present a Text-to-image Ambiguity Benchmark (TAB) dataset and a disambiguation framework for generating images that more closely align with user intention, as well as novel automatic evaluation procedures for assessing ambiguity resolution. They employ few-shot learning with specific language models to disambiguate ambiguous prompts with the aid of human feedback. For ***immoral issues***, Park *et al*. [133] first propose effective ethical image manipulation methods by localizing immoral attributes, including blurring, inpainting, and text-driven image manipulation, that have demonstrated effectiveness. For ***stereotypes issues***, Struppek *et al*. [168] demonstrate that text-based image genera-

tion models are sensitive to character encodings, with the insertion of even a single homoglyph at an arbitrary position can introduce cultural biases and stereotypes into the generated images, thereby influencing the image generation process. Additionally, Bansal *et al*. [7] indicate that certain keywords, such as 'irrespective of gender' and 'culture', can trigger substantial variations and diversities in model predictions, particularly in the context of gender bias within ethical interventions. Federico *et al*. [12] investigate accessible text-to-image generation models and expose the extent of categorization, stereotyping, and complex biases in the Stable Diffusion model [150] and generated images.

5.3. Challenges & Future Outlooks

🎯 As mentioned in Sec. 2.1, empirical evidence has demonstrated that supervised learning models are highly effective in accomplishing tasks for which they have been specifically trained by leveraging labeled data. However, their performance tends to decline when confronted with tasks beyond their range, as their proficiency is heavily dependent on the quality of the labeled data. Furthermore, it is impractical to label every piece of information available worldwide. Consequently, there has been a growing emphasis on developing more versatile, generalist models within the field of AI. The aim is to develop models capable of performing well across a range of tasks, thereby reducing dependence on vast quantities of labeled data.

It is noteworthy that the existing text-to-face datasets suffer from a lack of large-scale image-text pairs, which inadequacy significantly impedes progress in automatic text-to-face synthesis research. The primary reason lies in the burdensome process of collecting and annotating facial images. Furthermore, unlike other image categories, such as birds or flowers, facial features are more complex and multifaceted, including numerous factors such as ethnicity, gender, age, expression, and environmental context [223]. Therefore, developing large-scale text-to-face datasets entails greater challenges than those associated with other image domains.

🎯 As discussed in Sec. 4.1, text-to-face technology can potentially obtain valuable support from eyewitness testimony. However, the reliability of such testimony may be compromised due to factors such as fear or cognitive limitations, leading to inconsistencies between the provided description and the suspect's actual physical appearance. Therefore, the ability to manipulate specific visual features in synthetic facial images while maintaining other attributes consistent with the input description has become increasingly crucial in developing text-to-face synthesis technology for the public safety domain.

This necessitates an in-depth investigation aimed at improving the effectiveness of text-to-face synthesis approaches while adhering to the following specific compliance requirements: (1) *Discriminability*: Ensuring that the generated images are recognizable as individual persons. (2) *High Resolution*: Producing images with high resolution to facilitate detailed analysis. (3) *Photorealism*: Generating images that closely resemble authentic faces. (4) *Diversity*: Creating a variety of images from multiple viewpoints. (5) *Fidelity*: Ensuring that the generated images are consistent with the input description. (6) *Controllability*: Enabling selective manipulation of different attributes with text prompts while preserving other irrelevant attributes.

🎯 As mentioned in Sec. 4.2 and Sec. 4.3, text-to-X and X-to-image tasks are subcategories of multimodal learning. Here are some challenges: (1) *Alignment*: Ensuring that different modalities are appropriately aligned so that the generated results accurately reflect the input modality. (2) *Data scarcity*: Collecting and annotating large-scale multimodal datasets is time-consuming and expensive, particularly for specialized domains, which limit the performance of existing models. (3) *Scalability*: Developing models that efficiently handle large-scale multimodal data without compromising performance is an ongoing challenge. This includes addressing the memory and computational requirements associated with managing multiple modalities.

🎯 Despite the use of thoughtfully designed prompts to promote diversification and subvert undesired stereotypes, the limitations of text-to-image generation models are evident. As discussed in Sec. 5.1 and Sec. 5.2, several instances have illustrated the fragility of these models, despite their remarkable ability to generate images. This issue remains unsolved and requires significant further research.

Universal access and commercial use of open-source AI generators have been a noticeable trend, which may have both positive and negative consequences. Some may contend that regardless of its quality, AI technology should be accessible to everyone. Nevertheless, the long-term implications of this trend remain uncertain and require further evaluation and consideration.

6. Conclusion

Situated at the intersection of theoretical foundations and practical applications, multimodal learning has reached a critical juncture in today's rapidly evolving landscape. In recognition of the importance of synthesizing multiple domains, our objective in this review is to provide a comprehensive and scholarly analysis of contemporary advancements in multimodal learning. This survey aims to seamlessly connect theoretical underpinnings with real-world considerations by analyzing and summarizing various vision and language methodologies, innovative techniques, and emerging trends. Ultimately, this work seeks to build an integrated framework and highlight current challenges and opportunities that facilitate a deeper understanding of multimodal learning in the research community and its real-world implications.

References

[1] Jorge Agnese, Jonathan Herrera, Haicheng Tao, and Xingquan Zhu. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(4):e1345, 2020. 1

[2] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022. 6

[3] Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. Rhythm is a dancer: Music-driven motion synthesis with global structure. *arXiv preprint arXiv:2111.12159*, 2021. 6

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 5

[5] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019. 4

[6] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 4

[7] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022. 8

[8] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. 2023. 6

[9] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017. 3

[10] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 5

[11] Camilo Bermudez, Andrew J Plassard, Larry T Davis, Allen T Newton, Susan M Resnick, and Bennett A Landman. Learning implicit brain mri manifolds with deep learning. In *Medical Imaging 2018: Image Processing*, volume 10574, pages 408–414. SPIE, 2018. 5

[12] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*, 2022. 8

[13] Cristian Bodnar. Text to image synthesis using generative adversarial networks. *arXiv preprint arXiv:1805.00676*, 2018. 1

[14] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 4

[15] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 5

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 7

[17] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pages 185–200, 2018. 3

[18] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[19] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023. 1

[20] Miriam Cha, Youngjune Gwon, and HT Kung. Adversarial nets with perceptual losses for text-to-image synthesis. In *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2017. 4

[21] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 6

[22] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023. 4

[23] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 6

[24] Xiang Chen, Lingbo Qing, Xiaohai He, Xiaodong Luo, and Yining Xu. Ftgan: A fully-trained generative adversarial networks for text to face generation. *arXiv preprint arXiv:1904.05729*, 2019. 2, 4

[25] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10911–10920, 2020. 4

[26] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan2: Rich feature generation for text-to-image synthesis from constrained prior knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5187–5200, 2021. 4

[27] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022. 4

[28] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 5

[29] Thirza Dado, Yağmur Güçlütürk, Luca Ambrogioni, Gabriëlle Ras, Sander Bosch, Marcel van Gerven, and Umut Güçlü. Hyperrealistic neural decoding for reconstructing faces from fmri activations via the gan latent space. *Scientific reports*, 12(1):141, 2022. 5, 6

[30] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018. 3

[31] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Irc-gan: Introspective recurrent convolutional gan for text-to-video generation. In *IJCAI*, pages 2216–2222, 2019. 4

[32] 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. 4

[33] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 4

[34] Tan M Dinh, Rang Nguyen, and Binh-Son Hua. Tise: Bag of metrics for text-to-image synthesis evaluation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 594–609. Springer, 2022. 2, 3

[35] Hao Dong, Jingqing Zhang, Douglas McIlwraith, and Yike Guo. I2t2i: Learning text to image synthesis with textual data augmentation. In *2017 IEEE international conference on image processing (ICIP)*, pages 2015–2019. IEEE, 2017. 4

[36] Yanlong Dong, Ying Zhang, Lin Ma, Zhi Wang, and Jiebo Luo. Unsupervised text-to-image synthesis. *Pattern Recognition*, 110:107573, 2021. 4

[37] Xiaodan Du, Raymond A Yeh, Nicholas Kolkin, Eli Shechtman, and Greg Shakhnarovich. Text-free learning of a natural language interface for pretrained face generators. *arXiv preprint arXiv:2209.03953*, 2022. 4

[38] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10304–10312, 2019. 4

[39] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 5

[40] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. 6

[41] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022. 1

[42] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 4

[43] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *arXiv preprint arXiv:2210.15257*, 2022. 4

[44] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021. 4

[45] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209, 2021. 1

[46] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 717–734. Springer, 2022. 5

[47] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. *arXiv preprint arXiv:2211.12824*, 2022. 5

[48] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 4

[49] Philip Galanter. Computational aesthetic evaluation: past and future. *Computers and creativity*, pages 255–293, 2012. 6

[50] Albert Gatt, Marc Tanti, Adrian Muscat, Patrizia Paggio, Reuben A Farrugia, Claudia Borg, Kenneth P Camilleri, Mike Rosner, and Lonneke Van der Plas. Face2text: Collecting an annotated image description corpus for the generation of rich face descriptions. *arXiv preprint arXiv:1803.03827*, 2018. 2

[51] Subhankar Ghosh, Prasun Roy, Saumik Bhattacharya, Umapada Pal, and Michael Blumenstein. Tic: Text-guided

image colorization. *arXiv preprint arXiv:2208.02843*, 2022. 5

[52] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3, 6

[53] Satya Krishna Gorti and Jeremy Ma. Text-to-image-to-text translation using cycle consistent adversarial networks. *arXiv preprint arXiv:1808.04538*, 2018. 4

[54] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655*, 2023. 1

[55] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 4

[56] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 4

[57] Sonam Gupta, Arti Keshari, and Sukhendu Das. Rv-gan: Recurrent gan for unconditional video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2024–2033, 2022. 3

[58] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis. *arXiv preprint arXiv:2302.12764*, 2023. 6

[59] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022. 6

[60] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022. 7

[61] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15049–15058, 2021. 5

[62] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 4

[63] 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. 2

[64] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1552–1565, 2020. 2, 4

[65] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 4, 5

[66] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[67] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018. 4

[68] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 4

[69] Xianxu Hou, Xiaokang Zhang, Yudong Li, and Linlin Shen. Textface: Text-to-style mapping based face generation and manipulation. *IEEE Transactions on Multimedia*, 2022. 4

[70] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: Controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022. 5

[71] Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*, 2023. 5

[72] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 91–109. Springer, 2022. 6

[73] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1138–1147, 2021. 4

[74] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *arXiv preprint arXiv:2105.06458*, 2021. 5

[75] Nidhi Jain, Aria Wang, Margaret M Henderson, Ruogu Lin, Jacob S Prince, Michael J Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1):175, 2023. 5

[76] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5202, 2020. 3

[77] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial edit-

ing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021. 2

[78] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 2

[79] Bu Jin, Beiwen Tian, Hao Zhao, and Guyue Zhou. Language-guided semantic style transfer of 3d indoor scenes. *arXiv preprint arXiv:2208.07870*, 2022. 5

[80] KJ Joseph, Arghya Pal, Sailaja Rajanala, and Vineeth N Balasubramanian. C4synth: Cross-caption cycle-consistent text-to-image synthesis. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 358–366. IEEE, 2019. 4

[81] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. *arXiv preprint arXiv:2303.05511*, 2023. 3

[82] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. *arXiv preprint arXiv:2303.05511*, 2023. 6

[83] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *arXiv preprint arXiv:2210.02438*, 2022. 4

[84] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[85] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 4

[86] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[87] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020. 4

[88] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 5

[89] Taehoon Kim, Gwangmo Song, Sihaeng Lee, Sangyun Kim, Yewon Seo, Soonyoung Lee, Seung Hwan Kim, Honglak Lee, and Kyunghoon Bae. L-verse: Bidirectional generation between image and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16526–16536, 2022. 6

[90] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, 2006. 5

[91] Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. Speech2properties2gestures: Gesture-property prediction as a tool for generating representational gestures from speech. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 145–147, 2021. 6

[92] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 4

[93] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 5

[94] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 4

[95] Seung Hyun Lee, Chanyoung Kim, Wonmin Byeon, Gyeongrok Oh, Jooyoung Lee, Sang Ho Yoon, Jinkyu Kim, and Sangpil Kim. Robust sound-guided image manipulation. *arXiv preprint arXiv:2208.14114*, 2022. 5, 6

[96] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 3

[97] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[98] Jiguo Li, Xinfeng Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and Wen Gao. Direct speech-to-image translation. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):517–529, 2020. 6

[99] Ruijun Li, Weihua Li, Yi Yang, Hanyu Wei, Jianhua Jiang, and Quan Bai. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. *arXiv preprint arXiv:2210.09549*, 2022. 4

[100] Wei Li, Shiping Wen, Kaibo Shi, Yin Yang, and Tingwen Huang. Neural architecture search with a lightweight transformer for text-to-image synthesis. *IEEE Transactions on Network Science and Engineering*, 9(3):1567–1576, 2022. 4

[101] Wei Li, Xue Xu, Xinyan Xiao, Jiachen Liu, Hu Yang, Guohao Li, Zhanpeng Wang, Zhifan Feng, Qiaoqiao She, Yajuan Lyu, et al. Upainting: Unified text-to-image diffusion generation with cross-modal guidance. *arXiv preprint arXiv:2210.16031*, 2022. 3

[102] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019. 4

[103] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae

Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023. 4

[104] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 4

[105] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18197–18207, 2022. 4

[106] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 491–508. Springer, 2020. 4

[107] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18187–18196, 2022. 4

[108] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022. 3, 4

[109] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[110] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 4

[111] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 7

[112] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022. 7

[113] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Haoran Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021. 4

[114] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023. 6

[115] Xiaodong Luo, Xiang Chen, Xiaohai He, Linbo Qing, and Xinyue Tan. Cmafgan: A cross-modal attention fusion based generative adversarial network for attribute word-to-face synthesis. *Knowledge-Based Systems*, 255:109750, 2022. 4

[116] Xiaodong Luo, Xiaohai He, Xiang Chen, Linbo Qing, and Jin Zhang. Dualg-gan, a dual-channel generator based generative adversarial network for text-to-face synthesis. *Neural Networks*, 155:155–167, 2022. 4

[117] Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Is the elephant flying? resolving ambiguities in text-to-image generative models. *arXiv preprint arXiv:2211.12503*, 2022. 7

[118] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4

[119] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 4

[120] Vihaan Misra, Peter Schaldenbrand, and Jean Oh. Robot synesthesia: A sound and emotion guided ai painter. *arXiv preprint arXiv:2302.04850*, 2023. 5, 6

[121] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems*, 31, 2018. 5

[122] Minheng Ni, Chenfei Wu, Haoyang Huang, Daxin Jiang, Wangmeng Zuo, and Nan Duan. N\" uwa-lip: Language guided image inpainting with defect-free vqgan. *arXiv preprint arXiv:2202.05009*, 2022. 5

[123] 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

[124] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 2

[125] OpenAI. Gpt-4 technical report. 2023. 1

[126] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, pages 192–202, 2022. 6

[127] Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988*, 2022. 6

[128] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*, 2023. 4

[129] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 4

[130] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *NeurIPS Datasets and Benchmarks*, 2021. 1

[131] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 3

[132] Hyojin Park, Youngjoon Yoo, and Nojun Kwak. Mc-gan: Multi-conditional generative adversarial network for image synthesis. *arXiv preprint arXiv:1805.01123*, 2018. 4

[133] Seongbeom Park, Suhong Moon, and Jinkyu Kim. Judge, localize, and edit: Ensuring visual commonsense morality for text-to-image generation. *arXiv preprint arXiv:2212.03507*, 2022. 7

[134] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 5

[135] Nikita Pavlichenko and Dmitry Ustalov. Best prompts for text-to-image models and how to find them. *arXiv preprint arXiv:2209.11711*, 2022. 6

[136] Martin Pernuš, Clinton Fookes, Vitomir Štruc, and Simon Dobrišek. Fice: Text-conditioned fashion image editing with guided gan inversion. *arXiv preprint arXiv:2301.02110*, 2023. 5

[137] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022. 4

[138] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*, 2022. 3

[139] Justin NM Pinkney and Chuan Li. clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip. *arXiv preprint arXiv:2210.02347*, 2022. 4

[140] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 4, 5

[141] Victor R. Preedy and Ronald R. Watson, editors. *5-Point Likert Scale*, pages 4288–4288. Springer New York, New York, NY, 2010. 3

[142] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019. 4

[143] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. 2023. 4

[144] Aashish Anantha Ramakrishnan, Sharon X Huang, and Dongwon Lee. Anna: Abstractive text-to-image synthesis with filtered news captions. *arXiv preprint arXiv:2301.02160*, 2023. 2

[145] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 4, 7

[146] 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. 1, 4, 7

[147] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 4

[148] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Advances in neural information processing systems*, pages 217–225, 2016. 4

[149] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021. 7

[150] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 4, 6, 7, 8

[151] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022. 7

[152] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 4

[153] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[154] Mohamed Shawky Sabae, Mohamed Ahmed Dardir, Remonda Talaat Eskarous, and Mohamed Ramzy Ebbed. Stylet2f: Generating human faces from textual description using stylegan2. *arXiv preprint arXiv:2204.07924*, 2022. 4

[155] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3, 4

[156] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 2

[157] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. Styleclipdraw: Coupling content and style in text-to-drawing translation. *arXiv preprint arXiv:2202.12362*, 2022. 5

[158] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo

Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2

[159] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[160] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. Chatpainter: Improving text to image generation using dialogue. *arXiv preprint arXiv:1802.08216*, 2018. 4

[161] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in computational neuroscience*, 13:21, 2019. 5

[162] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. 5

[163] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 4

[164] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 4

[165] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 6

[166] Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, pages 87–96. PMLR, 2017. 3

[167] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3, 4

[168] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022. 7

[169] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. Anyface: Free-style text-to-face synthesis and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18687–18696, 2022. 4

[170] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. Multi-caption text-to-face synthesis: Dataset and algorithm. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2290–2298, 2021. 4

[171] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, pages 2022–11, 2022. 5, 6

[172] Hongchen Tan, Xiuping Liu, Xin Li, Yi Zhang, and Baocai Yin. Semantics-enhanced adversarial nets for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10501–10510, 2019. 4

[173] Hongchen Tan, Xiuping Liu, Meng Liu, Baocai Yin, and Xin Li. Kt-gan: Knowledge-transfer generative adversarial network for text-to-image synthesis. *IEEE Transactions on Image Processing*, 30:1275–1290, 2020. 4

[174] Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. Dr-gan: Distribution regularization for text-to-image generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 4

[175] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 4, 5

[176] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 3

[177] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[178] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 4

[179] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[180] Bo Wang, Tao Wu, Minfeng Zhu, and Peng Du. Interactive image synthesis with panoptic layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7783–7792, 2022. 5

[181] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. 1

[182] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 4

[183] Hao Wang, Guosheng Lin, Steven C. H. Hoi, and Chunyan Miao. Cycle-consistent inverse gan for text-to-image synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 630–638, 2021. 4

[184] Min Wang, Congyan Lang, Liqian Liang, Songhe Feng, Tao Wang, and Yutong Gao. End-to-end text-to-image synthesis with spatial constrains. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–19, 2020. 4

[185] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 6

[186] Tianren Wang, Teng Zhang, and Brian Lovell. Faces a la carte: Text-to-face generation via attribute disentanglement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3380–3388, 2021. 2, 4

[187] Xinsheng Wang, Tingting Qiao, Jihua Zhu, Alan Hanjalic, and Odette Scharenborg. Generating images from spoken descriptions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:850–865, 2021. 5, 6

[188] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022. 4

[189] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 2

[190] Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *arXiv preprint arXiv:2207.09814*, 2022. 6

[191] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022. 6

[192] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 6

[193] Fuxiang Wu, Liu Liu, Fusheng Hao, Fengxiang He, and Jun Cheng. Text-to-image synthesis based on object-guided joint-decoding transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18113–18122, 2022. 4

[194] Lin Wu, Yang Wang, Feng Zheng, Qi Tian, and Meng Wang. T-person-gan: Text-to-person image generation with identity-consistency and manifold mix-up. *arXiv preprint arXiv:2208.12752*, 2022. 4

[195] Xianchao Wu. Creative painting with latent diffusion models. *arXiv preprint arXiv:2209.14697*, 2022. 4

[196] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4

[197] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems*, 34:2598–2610, 2021. 3

[198] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022. 5

[199] 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. 2, 4

[200] Zhen Xu, Si Wu, Qianfen Jiao, and Hau-San Wong. Tsev-gan: Generative adversarial networks with target-aware style encoding and verification for facial makeup transfer. *Knowledge-Based Systems*, 257:109958, 2022. 3

[201] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao. Elegant: Exquisite and locally editable gan for makeup transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 737–754. Springer, 2022. 3

[202] Yanhua Yang, Lei Wang, De Xie, Cheng Deng, and Dacheng Tao. Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis. *IEEE Transactions on Image Processing*, 30:2798–2809, 2021. 4

[203] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7764–7773, 2022. 5

[204] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2327–2336, 2019. 4

[205] 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*, 2022. 3, 4

[206] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021. 1

[207] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, and Shijian Lu. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021. 1

[208] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 1

[209] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3, 4

[210] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 4

[211] Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *arXiv preprint arXiv:2112.15283*, 2021. 6

[212] Junzhe Zhang, Daxuan Ren, Zhongang Cai, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Monocular 3d object reconstruction with gan inversion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 673–689. Springer, 2022. 3

[213] Kuan Zhang, Haoji Hu, Kenneth Philbrick, Gian Marco Conte, Joseph D Sobek, Pouria Rouzrokh, and Bradley J Erickson. Soup-gan: Super-resolution mri using generative adversarial networks. *Tomography*, 8(2):905–919, 2022. 3

[214] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 4

[215] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 4

[216] Xian Zhang, Xin Wang, Canghong Shi, Zhe Yan, Xiaojie Li, Bin Kong, Siwei Lyu, Bin Zhu, Jiancheng Lv, Youbing Yin, et al. De-gan: Domain embedded gan for high quality face image inpainting. *Pattern Recognition*, 124:108415, 2022. 3

[217] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. Ufc-bert: Unifying multi-modal controls for conditional image synthesis. *Advances in Neural Information Processing Systems*, 34:27196–27208, 2021. 2, 6

[218] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6199–6208, 2018. 4

[219] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 4

[220] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 7

[221] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 7

[222] Rui Zhou, Cong Jiang, and Qingyang Xu. A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing*, 451:316–336, 2021. 1

[223] Yutong Zhou. Generative adversarial network for text-to-face synthesis and manipulation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2940–2944, 2021. 2, 4, 8

[224] Yutong Zhou and Nobutaka Shimada. Generative adversarial network for text-to-face synthesis and manipulation with pretrained bert model. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 4

[225] Yutong Zhou and Nobutaka Shimada. Able: Aesthetic box lunch editing. In *Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and related APPlications*, pages 53–56, 2022. 2

[226] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021. 4

[227] Bin Zhu and Chong-Wah Ngo. Cookgan: Causality based text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5519–5527, 2020. 4

[228] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019. 4

[229] Yiming Zhu, Hongyu Liu, Yibing Song, Xintong Han, Chun Yuan, Qifeng Chen, Jue Wang, et al. One model to edit them all: Free-form text-driven image manipulation with semantic modulations. *arXiv preprint arXiv:2210.07883*, 2022. 5