RobotDesign1M: A Large-scale Dataset for Robot Design Understanding

Tri Le[1], Toan Nguyen[1], Quang Tran[2], Quang Nguyen[1], Baoru Huang[2], Hoan Nguyen[3], Minh Nhat Vu[4], Tung D. Ta[5,6], Anh Nguyen[2]

https://airvlab.github.io/robotdesign1m/

***Abstract*— Robot design is a complex and time-consuming process that requires specialized human expertise. Gaining a deeper understanding of robot design data can enable various applications such as automated design generation, retrieving example designs from the text prompt, or developing AI design assistants. While recent advancements in foundation models present promising approaches to addressing these challenges, progress in this field is hindered by the lack of large-scale robot design datasets. In this paper, we introduce RobotDesign1M, a large-scale dataset with 1 million samples. Our dataset features multimodal data collected from scientific literature, covering various robotics domains. We propose a semi-automated data collection pipeline, enabling efficient and diverse data acquisition. To assess the effectiveness of our new RobotDesign1M dataset, we conduct extensive experiments across multiple tasks, including design image generation, visual question answering about designs, and design image retrieval. The experimental results demonstrate that RobotDesign1M can be used as a challenging benchmark for several robot design understanding tasks. Furthermore, the cross-dataset evaluations demonstrate the ability of RobotDesign1M in improving model generalization, proving its data quality.**

I. Introduction

Foundation models such as ChatGPT trained on large-scale datasets have demonstrated significant potential across various robotics applications [1], [2]. Recently, several studies have explored the use of large foundation models to support the robot design process, aiming to lower costs, reduce human effort, and push beyond the limits of human creativity [2]–[9]. Despite these advancements, the ability of foundation models to deeply understand and contribute to robot design remains largely underexplored. Designing robots is a highly complex task that integrates knowledge from multiple disciplines and involves a series of critical steps, from functional requirements to physical implementation [3], [10]. Developing foundation models specifically for robot design has the potential to transform the field by leveraging intelligent, data-driven insights to enhance creativity, efficiency, and performance. In practice, those models can offer insightful feedback, suggest improvements, streamline workflows, and drive innovation [11].

While foundation models are useful to support the robot design process, the lack of large-scale robot design datasets is a major problem [12], [13]. Efforts have been made to expand datasets for other engineering domains to fuel modern deep learning models [14]–[16]. Most of these datasets are synthesized by trying and testing generated parameters through simulation [14], [17]. Other datasets [16], [18] provide computer-aided design (CAD) to facilitate research in engineering design generation. Recently, Ghezelbash *et al.* [15] proposes a small-scale dataset comprising mechanical designs and text descriptions. Unlike other generic engineering domains, robot design involves not only the geometric properties of individual components but also their attributes and interrelationships [19]. The absence of specialized large-scale robotic design datasets limits the development of dedicated foundation models tailored for robot design.

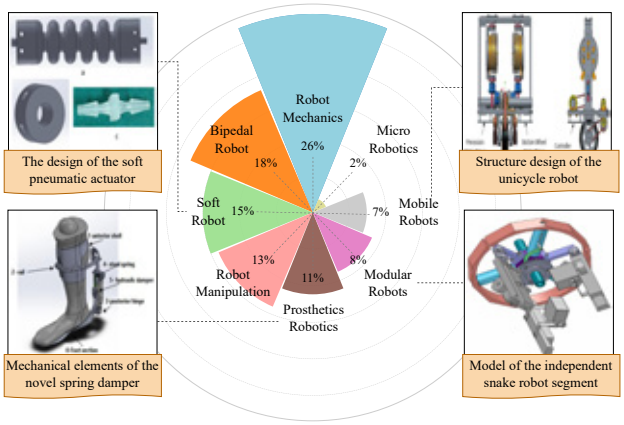


Fig. 1. We introduce RobotDesign1M, a new large-scale dataset with 1M samples covering various designs from different robotic disciplines.

In practice, collecting a large-scale robot design dataset is a non-trivial task and a time-consuming process. First, collecting technical design data is challenging as robot design involves highly specialized knowledge, and it cannot be generated by existing generative models (e.g., text to images) alone as in [30]. While datasets for robot design do exist, they are often proprietary and restricted within the industry, making them inaccessible at scale for the community [17]. Second, collecting large-scale robot design data requires an automated data collection pipeline and a strategy to support the labeling process. In this work, we address these two fundamental challenges. We first propose collecting robot designs from published scientific literature. These sources encompass a wide variety of technical designs of different robotic domains, reflecting extensive research efforts of the community over time. Then, we propose a data creation pipeline with a semi-automated data filtering and labeling process. In addition, we utilize a Large Language Model (LLM) to automatically construct visual instruction-following data, enhancing the utility of the curated data for recent advancements in multi-modal learning [31]–[35].

[1] FPT Software AI Center, Vietnam
[2] University of Liverpool, UK
[3] University of Information Technology, HCMC, Vietnam
[4] Automation & Control Institute, TU Wien, Austria
[5] Department of Creative Informatics, The University of Tokyo, Japan
[6] Faculty of Environment and Information Studies, Keio University, Japan

TABLE I

COMPARISON OF DESIGN DATASETS.

Dataset	#Samples	Design Representation	Domain	Caption	Conversation	Real Design	Public
Daele *et al.* [20]	5K	Technical Drawing	Generic	✘	✘	✔	✘
Khan *et al.* [21]	400	Technical Drawing	Engineering	✘	✘	✔	✘
ABC [22]	1M	CAD Model	Generic	✘	✘	✔	✔
DeepCAD [23]	178K	CAD Model	Generic	✘	✘	✔	✔
Omni-CAD [16]	453K	CAD Model	Generic	✔	✘	✔	✔
CAD Fusion [24]	20K	CAD Model	Generic	✔	✘	✔	✔
Text2CAD [18]	170K	CAD Model	Generic	✔	✘	✔	✔
Fusion 360 Gallery [25]	8K	CAD Model	Generic	✘	✘	✔	✔
SketchGraph [26]	15M	CAD Sketch	Generic	✘	✘	✔	✔
Khan *et al.* [27]	300	CAD Image	Generic	✘	✘	✔	✘
LINKS [14]	100M	Parameter	Engineering	✘	✘	✘	✔
Ghezelbash *et al.* [15]	9K	CAD Image	Engineering	✔	✘	✔	✔
Jayanti *et al.* [28]	867	CAD Model	Engineering	✘	✘	✔	✔
Lee *et al.* [29]	715K	CAD Model	Engineering	✘	✘	✘	✔
RobotDesign1M (ours)	∼ 1M	CAD Image Technical Drawing	Robotics	✔	✔	✔	✔

We introduce **RobotDesign1M**, a large-scale dataset for robot design with approximately 1 million (1M) samples. In contrast to prior datasets [15], [18], [20], [22]–[24], [36], our dataset consists of 2D drawings and images of robot design and offers two key advantages: *(i)* it is a dedicated dataset for robot design, covering various types of robots and design aspects; *(ii)* it provides high-reliability figures and text sourced from scientific documents. Fig. 1 illustrates the robot design topics distribution of the collected scientific documents, along with representative data samples from selected topics. We empirically demonstrate that RobotDesign1M enhances the robot design understanding of general-domain models across multiple tasks, including visual question answering, text-image retrieval, and text-to-design image generation. Our findings confirm that general-domain models finetuned on RobotDesign1M outperform those finetuned on other related datasets. In summary, our contributions are as follows:

- We introduce RobotDesign1M, a large-scale dataset dedicated to robot design with over 1M samples and multimodal ground truth.
- We intensively benchmark several models on various design-related tasks with RobotDesign1M, proving that the quality and diversity of our dataset improve model generalization and validate it as a challenging benchmark for robot design understanding.

II. RELATED WORK

Datasets for Robot Design. Robotics is inherently interdisciplinary, integrating insights from several fields [37]. Consequently, a comprehensive robot design dataset must draw from multiple domains [38]. We review existing robot-related design datasets in Table I. Most current design datasets [16], [18], [20], [22]–[27], [36], [39] predominantly target generic objects. For example, ABC [22] is a large collection of CAD models in B-rep format and has served as a foundation for later datasets such as DeepCAD [23], Omni-CAD [16], Text2CAD [18]. Although extensive and diverse, these datasets only contain CAD models of generic objects. In contrast, recent works have begun to address the specific needs of mechanical engineering. Lee *et al.* [29] introduce a large dataset of 3D CAD models featuring mechanical parts. Ghezelbash *et al.* [15] propose a dataset dedicated to mechanisms and gears. However, these datasets remain confined to mechanical engineering, leaving a gap in interdisciplinary robotic domains. Our RobotDesign1M fills this gap by collecting large-scale data from multiple robotic fields, for example, soft robots, bipedal robots, etc.

Robot Design Automation. Numerous approaches have been studied for automating robot design process [3], such as evolutionary algorithms [1], [5], reinforcement learning [6], [40], and topology optimization [7]. Recently, generative models have been leveraged for robot design automation [4], [8], [9]. Wang *et al.* [4] introduce a framework that integrates diffusion models with physics simulation for morphology co-design. Song *et al.* [41] utilize an LLM within an evolutionary design loop to generate soft robot designs. Stella *et al.* [11] leverage LLM to discuss ideas and outline the specifications for the robot design. Jadhav *et al.* [42] combine LLM with the finite element method to assess and refine design outcomes. These early studies of LLM applications in robot design inspire our investigation into a potential research direction: creating a large-scale dataset that benefits the fine-tuning process for robot design. Distinct from existing robot design datasets [1], [5], [9], which comprise parameter configurations for specific robot designs, our RobotDesign1M dataset includes textual descriptions, 2D drawings, and visual instruction data to advance robot design understanding.

Multi-modal Learning. Recently, multi-modal learning has been widely applied in several tasks [43], [44], owing to the significant progress in learning [32], [34], [35], [45] and the availability of multi-modal datasets [46], [47]. In robot design, several multi-modal learning models are useful when applied in the design process. Text-image retrieval supports design knowledge acquisition, comparative analysis of designs, and inspiration discovery [13], [48]. Visual question answering assists engineers in analyzing design choices, providing useful feedback [11], and offers students

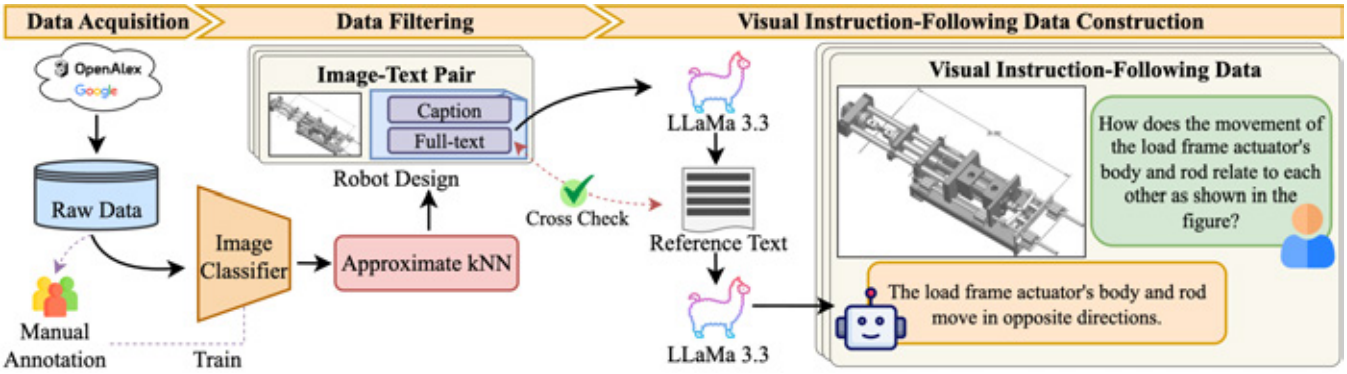


Fig. 2. **Dataset creation pipeline.** Our semi-automated dataset creation pipeline comprises three stages: raw data acquisition, data filtering, and visual instruction-following data construction.

guidance in their curriculum [49]. In recognizing these useful applications, we integrate LLM in our pipeline to construct the data that supports multi-modal learning frameworks [35].

III. The RobotDesign1M Dataset

Fig. 2 presents an overview of our dataset creation pipeline. We first collect scientific works from the public domain, followed by extracting images and text from these documents. To ensure relevance, we apply a series of semi-automated filtering processes to refine the extracted relevant data. Finally, these images and texts are utilized to construct an instruction-following dataset for further finetuning.

A. Data Acquisition

We collect scientific documents from multiple public domain sources, including Google Search, OpenAlex [50], and open access of IEEE Xplore [51]. To systematically retrieve robot design documents, we first compile a list of robotic keywords extracted from IEEE Robotics & Automation Letters [52]. This list is expanded using ChatGPT [53], which generates semantically similar terms and conceptually related keywords. After manual filtering, we finalize a set of 1K keywords, which are then used to query our data sources. Beyond robotics, we further collect scientific publications from related domains, including mechanical engineering and electrical engineering [37], [54], [55]. This data acquisition process results in a corpus of more than 1M documents, covering diverse publication types, including conference proceedings, journal articles, and dissertations.

For content extraction, we utilize PDFFigures [56] to extract images, associated captions, and the full text from the collected documents. This yields a total of 5M image-text pairs, where each image is linked to a caption and the full text of its respective document.

B. Data Filtering

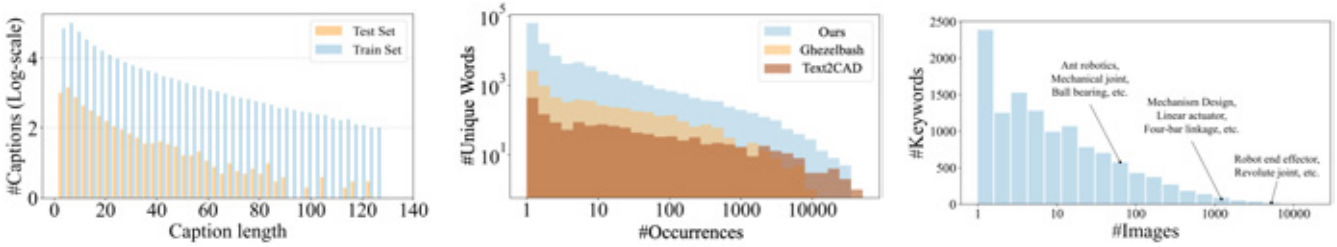
Irrelevant Data Filtering. The initial extraction includes noisy image-text pairs that are not related to robot design, such as charts, tables, and human-robot experiment images. To filter irrelevant images while maintaining a scalable and automated data collection pipeline, we train an image classifier that is an ensemble model comprising pretrained OpenCLIP [45] and pretrained EfficientNetV2 [57] as backbones. For the OpenCLIP model, we select the pretrained checkpoints with the highest performance on the ImageNet Sketch benchmark [58], as this benchmark has the most similar visual concept to robot design schematics. To train the classifier, we *manually annotate* 32,000 images, which are a subset of the extracted data. The classifier achieves over 95% accuracy on our held-out test set before being used to classify the full data. Despite a potential modest amount of noise in our dataset, we demonstrate that its high quality still greatly benefits downstream tasks in our experiments.

Deduplication. To remove duplicate samples, we use an approximate k-Nearest-Neighbor (kNN) method as in [59]. Specifically, we compute the embeddings of images using a pretrained OpenCLIP [45] and identify similar images using approximate kNN, implemented by Faiss [60]. Many robot designs are different despite the visual representations being similar. To prevent the erroneous removal of distinct designs, we first identify similar images using approximate kNN. We then compare their associated text descriptions to verify whether they represent the same design. Only samples with both highly similar images and nearly identical textual descriptions are removed. After completing the data filtering process, the final dataset contains approximately 1M image-text pairs relevant to robot design.

C. Visual Instruction-Following Data

Visual instruction-following data plays a crucial role in finetuning multi-modal learning models [34], [61]. However, such data remains scarce in the context of robot design, hindering the development of domain-specific multimodal models. To bridge this gap, we further leverage our curated data to construct a dedicated visual instruction-following dataset for robot design. In this data, each image is associated with a single or multiple question-answer pairs between User and Assistant, which interpret the visual content of the image.

Our visual instruction-following data construction process is designed to enhance reliability while ensuring scalability. First, we observe that several captions provide insufficient descriptions for the images. To enrich the textual content, we leverage an LLM to extract reference texts that are sentences mentioning the image in its respective document. To mitigate LLM hallucination [62], the extracted reference



(a) Caption length distribution. (b) Caption vocabulary. (c) Image keyword distribution.

Fig. 3. **Dataset Statistics**. We provide statistics on (a) caption length, (b) caption vocabulary, and (c) image keywords.

texts are validated by cross-checking them with the full text. Since LLMs are less prone to hallucination when processing short input [62], we prompt the LLM to generate question-answer pairs based on the caption and validated reference texts. We invite 5 robotic engineers to manually verify 100 generations before processing the remaining data. In the implementation, we experiment with GPT-4o [53], LLaMa 3.3 70B [63], and Qwen2.5 72B [64] and choose LLaMa 3.3 70B as our primary LLM, owing to the balance between affordability and the quality of generation. In addition, to ensure the reference text is extracted accurately, we only provide LLM with paragraphs that contain the image number. Fig. 4 illustrates an example of the caption, reference text, and generated question-answer pairs. We show more samples from our dataset in Fig. 5. Overall, we construct approximately 1.3 million question-answer pairs associated with the design images.

Caption: A cutaway view of the finger, showing the tendon channel of the proximal phalange.

Reference Text:
The key advantage of this joint design over the previous one is that it allows for the tendons to pass through the center of rotation of the joint, which will eliminate any coupling effect between the consecutive MP joint and the PIP- DIP joints.

Visual-Instruction Following Data:

User: What is the primary advantage of the joint design shown in the figure, in terms of tendon placement and joint rotation?
Assistant: The joint design allows for the tendons to pass through the center of rotation of the joint.
User: How does the tendon placement in the figure affect the interaction between consecutive joints?
Assistant: The tendon placement eliminates any coupling effect between the consecutive MP joint and the PIP-DIP joints.

Fig. 4. **Visual Instruction-Following Data.** Caption and Reference Text are extracted from the documents, while question-answer pairs are generated using LLaMa 3.3 70B [63].

D. RobotDesign1M Statistics

Caption Length Distribution. RobotDesign1M contains captions of varying lengths. Fig. 3(a) illustrates the distribution of caption lengths in the training and test sets. The average caption length across the dataset is 15.14 words.

Caption Vocabulary. To evaluate caption diversity, we apply lemmatization [65] to map words to their base forms based on context, resulting in a refined caption vocabulary. Fig. 3(b) compares the vocabulary sizes of RobotDesign1M, Ghezelbash *et al.* [15], and Text2CAD [18]. RobotDesign1M contains 113K unique words, excluding stop words and numbers. Our dataset vocabulary size is approximately 16-78 times higher than Ghezelbash [15] and Text2CAD [18].

Visual Content Analysis. In addition to textual diversity, we assess our dataset's visual diversity by analyzing the topics and keywords associated with each image. We assign each image the primary topics and keywords from its respective documents, as indexed by OpenAlex [50]. Fig. 1 presents the distribution of robot design topics that contain "robot" in their name and their description includes robot design aspects. We shorten the topic names for ease of presentation. The images in RobotDesign1M cover 1K topics classified by OpenAlex [50]. While topics provide a higher-level abstraction, we further examine image keywords to gain a more granular understanding of content diversity. Fig. 3(c) visualizes the distribution of image keywords. The images in RobotDesign1M are associated with over 12K keywords about robotics and related disciplines, covering a broad range of topics, such as mechanical joints, ball bearings, etc.

E. How will RobotDesign1M be used in the community?

We anticipate several promising directions that can benefit from our dataset. Two direct applications are:

- *Robot Design Chatbot*: Finetuning foundation models on visual instruction-following data enables intelligent interaction for robot design. The texts and images in RobotDesign1M are sourced from scientific works, ensuring accurate information for chatbot training.
- *Cross-modal Design Search*: Cross-modal design search assists engineers and students in looking for design by different modalities such as text and images, facilitating design analogy and inspiration exploration [13], [48], [66]. By providing captions and enriched, detailed texts associated with each design, RobotDesign1M allows training models that accept various types of queries, e.g., design name, functionality, and appearance.

Furthermore, with its large-scale nature, our RobotDesign1M dataset can be utilized to develop high-quality data filters [67], [68], robot design generation [2], [4], design retrieval [13], [48], and scientific image captioning [69].

IV. Experiments

To assess the utility of RobotDesign1M, we evaluate how models trained on it perform in comparison with those on related datasets on three robot design tasks: Visual question answering on robot design data; Retrieving design images from text; and Generating new designs from user prompts.

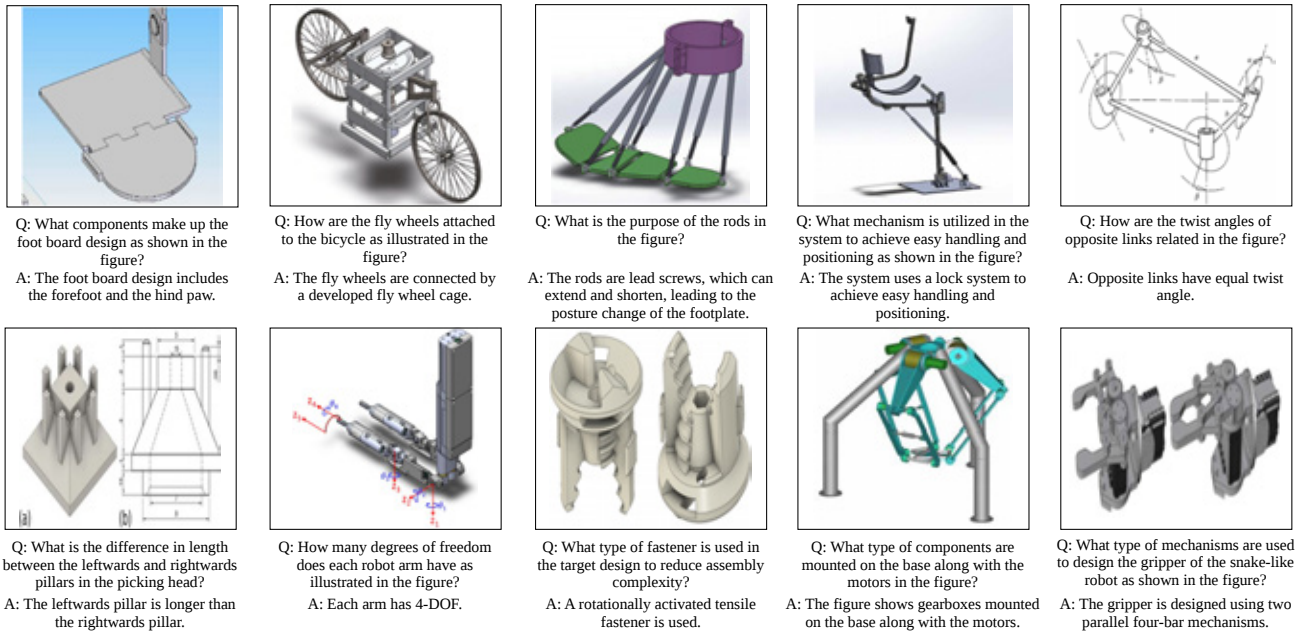


Fig. 5. **Samples for VQA task from RobotDesign1M.**

A. *Visual Question Answering in Robot Design*

Setup. In the visual question answering (VQA) task, a model is provided with an image and a textual question and is expected to generate a correct answer. To comprehensively assess the model's understanding of robot design, we construct open-ended question-answer pairs. We use common text generation evaluation metrics [70], specifically BLEU [71], METEOR [72], and L3Score [73]. For L3Score, we use GPT-4o as the evaluator as in [73].

Implementation. We finetune two pretrained LLM models: LlaVa-1.5 [74] and Qwen2-VL [75] on three datasets: RobotDesign1M, Text2CAD [18], and Ghezelbash *et al.* [15] (Ghezelbash). While Text2CAD and Ghezelbash are not specifically designed for robot design, they are the most comparable datasets, as their publicly available data modalities align closely with ours, and their domains focus on engineering design, which shares more similarities with robot design. For Text2CAD, we construct image-text pairs by utilizing multi-view images along with their corresponding generated textual descriptions [18]. Since both Text2CAD and Ghezelbash datasets include only image-text pairs, we adapt our pipeline to construct visual instruction-following data for a fair comparison. All models are trained on 4 NVIDIA A100-80GB GPUs until convergence.

TABLE II

OPEN-ENDED VISUAL QUESTION ANSWERING RESULTS

		RobotDesign1M (Ours)			**Text2CAD** [18]			**Ghezelbash** [15]		
Models	FT?	B↑	M↑	L↑	B↑	M↑	L↑	B↑	M↑	L↑
GPT-4o	No	**2.51**	**20.83**	0.30	3.12	7.74	0.30	**5.32**	11.52	0.38
LlaVa-1.5 [61]	No	0.82	4.63	0.12	3.06	6.69	0.27	1.90	7.22	0.20
Qwen2-VL [34]	No	0.61	6.23	0.13	1.63	7.08	0.30	1.32	6.83	0.27
LlaVa-1.5 [61]	Yes	0.83	9.80	0.18	6.75	**32.64**	0.94	3.54	17.90	0.48
Qwen2-VL [34]	Yes	1.32	12.33	**0.31**	**6.94**	**32.64**	**0.94**	3.62	**18.50**	**0.52**

VQA Results. Table II presents the BLEU (B), METEOR (M), and L3Score (L) of the finetuned (FT) models, along with the performance of original pretrained models (without finetuning) and GPT-4o as reference points. The results reveal two key observations. First, general LLMs without finetuning perform poorly across three datasets, highlighting a significant gap in their applicability to engineering and robotic design. While finetuning on domain-specific datasets improves performance, there remains considerable room for further enhancement. Second, Table II shows that our RobotDesign1M dataset is a *challenging benchmark as models finetuned on our dataset achieve lower accuracy in all metrics, compared to those trained on other datasets*. The low accuracy performance of all models in Table II also indicates that the current methods do not handle well the challenging cases in our dataset. However, we note that *models trained on RobotDesign1M are neither overfitting to the training set nor suffering from dataset noise*, as evidenced by their superior performance compared to models trained on other datasets in the cross-dataset evaluation (Table III).

TABLE III

CROSS-DATASET VQA GENERALIZATION RESULTS

Train \ Test	**Text2CAD** [18]			**Ghezelbash** [15]			**RobotDesign1M (Ours)**		
	B↑	M↑	L↑	B↑	M↑	L↑	B↑	M↑	L↑
Text2CAD [18]	6.94	32.64	0.94	1.32	8.84	0.12	0.44	6.50	0.09
Ghezelbash [15]	2.63	15.93	0.34	3.62	18.50	0.52	**0.84**	**9.93**	**0.19**
RobotDesign1M (Ours)	**3.03**	**17.42**	**0.47**	**2.12**	**12.93**	**0.33**	1.32	12.33	0.31

Cross-dataset VQA Generalization. Table III presents the results of finetuning LLMs on a dataset (row) and testing on another dataset (column). For example, the pretrained Qwen2-VL model finetuned on Text2CAD achieves a BLEU score of 0.013 when tested on Ghezelbash. We highlight the best metrics in cross-dataset evaluations in **bold** and mark in-dataset results with underlines. RobotDesign1M leads to an improvement of $\approx$ 15-62% compared to other datasets in the cross-dataset setting. *This proves the quality of data in RobotDesign1M in promoting model generalization.* Fig. 6 provides two qualitative examples from the RobotDesign1M test set. In these examples, models finetuned on the

Text2CAD and Ghezelbash datasets fail to correctly answer the questions for the given images.

User: What type of joints are used in the 3-UU system as shown in the figure?

Text2CAD: The 3-UU system has four cylindrical joints.
Ghezelbash: The 3-UU system uses only revolution joints.
Ours: The 3-UU system uses universal joints.

User: What is the actuation type of the robotic mechanism shown in the figure?

Text2CAD: The mechanical device is articulated with two intersecting cylindrical joints.
Ghezelbash: The actuation type is hydraulic.
Ours: The robotic mechanism is redundantly actuated.

Fig. 6. **Qualitative VQA Results on Test set.** Answers of Qwen2-VL finetuned on Text2CAD [18], Ghezelbash [15], and our RobotDesign1M. Correct answers are highlighted in green, while incorrect ones are in red.

B. Retrieving Design Images from Text

Setup. In text-image retrieval for robot design tasks, a model receives a text query and is tasked with retrieving robot design images that best match the query. We use Recall@k (R@k) [76] as the evaluation metric, which measures the proportion of ground-truth instances that appear within the top k retrieved results across all test cases. We report results for k values of 1, 5, and 10.

Implementation. We finetune a state-of-the-art multimodal encoder BLIP-2 [32] on three datasets: RobotDesign1M, Text2CAD [18], and Ghezelbash [15]. We leverage the image-caption pairs from the selected datasets as training and test data. In this setting, each text query corresponds to a single ground-truth image. All models are trained on 4 NVIDIA A100-80GB GPUs until convergence.

TABLE IV
TEXT-IMAGE RETRIEVAL RESULTS

Train \ Test	Text2CAD [18]			Ghezelbash [15]			RobotDesign1M (Ours)		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Text2CAD [18]	16.21	40.92	53.45	6.44	14.00	19.89	**5.36**	**10.10**	**12.98**
Ghezelbash [15]	1.01	5.04	8.74	22.11	49.44	63.00	4.78	9.86	13.20
RobotDesign1M (Ours)	**1.68**	**6.30**	**9.33**	**8.77**	**21.11**	**29.22**	12.9	27.46	36.28

Retrieval Results. Table IV presents the finetuning results of the BLIP-2 model on one dataset (rows) and evaluating it on another dataset (columns). Results suggest that RobotDesign1M is a challenging benchmark, as models finetuned on it achieve lower performance in all metrics, yet generalize better in cross-dataset testing, compared to those trained on other datasets. Fig. 7 shows top-1 retrieval results for queries from three test sets. The qualitative results show that the model finetuned on RobotDesign1M performs well across various design types, including CAD images and drawings.

C. Text-To-Design Image Generation

Setup. In this task, a model takes a textual description of the robot design as input and generates a corresponding design image. We use the FID [77] metric, a widely used metric for evaluating text-to-image generation quality [78]. FID is used to measure the distance between two distributions of generated images and the source distribution of the ground-truth data. The lower FID is better.

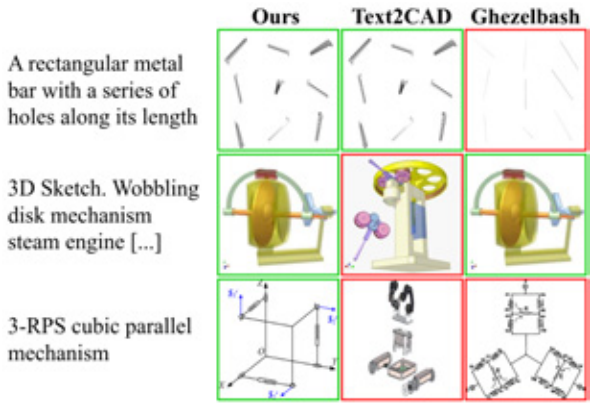


Fig. 7. **Top-1 Text-Image Retrieval Results**. The top-1 retrieval results of models finetuned on three datasets (columns) on three test sets: Text2CAD [18] (top), Ghezelbash [15] (middle), and RobotDesign1M (bottom). Correct samples are in green, and incorrect ones are in red boxes.

Implementation. We finetune a pretrained state-of-the-art image generation model, Stable Diffusion XL [79] on image-caption pairs of RobotDesign1M. We train the model on a server with 4 NVIDIA A100-80GB for 6 days with a batch size of 4 and a gradient accumulation of 8.

Generation Results. We report generation results in Table V. The results show that the model finetuned on RobotDesign1M significantly optimizes the FID metric. Fig. 8 presents images generated by GPT-4o, a pretrained Stable Diffusion XL model, and the finetuned model on our RobotDesign1M. The finetuned model produces more realistic images, while those generated from GPT-4o and the pretrained Stable Diffusion XL focus more on aesthetics.

TABLE V
TEXT-TO-IMAGE GENERATION RESULTS

Model	Finetuned on RobotDesign1M?	FID ↓
Stable Diffusion XL [79]	No	45.83
Stable Diffusion XL [79]	Yes	**39.42**

User Study Results. We conducted a user study with 12 robotics engineers and 1,200 evaluations on robot designs generated by GPT-4o, Stable Diffusion XL with and without finetuning on RobotDesign1M. Participants rated each generated image on a 5-point Likert scale based on its alignment with the input text. Fig. 9 presents the results. The model finetuned on RobotDesign1M received higher ratings compared to both the original pretrained model and GPT-4o.

D. Discussion

Through the experiments, we demonstrate the effectiveness of our RobotDesign1M in supporting robot design tasks. The diversity of our dataset across various robotic domains enhances the generalization capabilities of multimodal LLMs, as evidenced by its superior performance compared to other datasets in cross-dataset test settings (Table III and Table IV). While our dataset improves multimodal LLMs in the robotic domain (Table II and Table V), their suboptimal performance highlights the need for further advancements in model architectures to better capture the underlying patterns in robot design data.

Limitations. Although RobotDesign1M shows promising

Fig. 8. **Qualitative Results**. Images generated from the same prompts (column) by GPT-4o (top), the original Stable Diffusion XL (middle), and a finetuned version on RobotDesign1M (bottom). The model finetuned on our dataset produces more realistic design images.

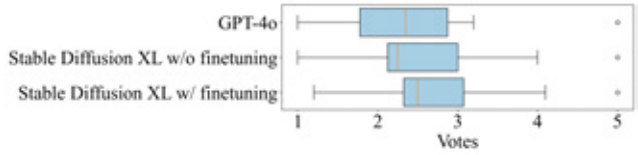


Fig. 9. **User Study Results**. Users voted on the design images generated by Stable Diffusion XL [79] with and without finetuning on our dataset.

results, there are limitations necessitating further studies. There is a trade-off between reliability and the richness of textual descriptions. In our dataset, we prioritize reliability and only provide the LLMs with short paragraphs explicitly referring to the images, thereby mitigating hallucinations [62]. However, this approach inadvertently excludes relevant text that implicitly describes the images. Developing methods to address hallucinations while increasing the richness of extracted texts would enhance the dataset's utility. In the future, incorporating additional modalities such as parametric designs, CAD models, or converting 2D images into 3D representations [47], [80] would provide a more comprehensive understanding of robot design, increasing the usefulness of our dataset.

V. Conclusions

We introduce **RobotDesign1M**, a new large-scale dataset for robot design understanding. Our dataset with 1M samples is constructed by a semi-automated data creation pipeline. RobotDesign1M possesses two unique attributes, which are its diversity over a broad range of robotic domains and its high reliability, owing to its data collection sources. The intensive experiments show the usefulness of RobotDesign1M as foundation models gain consistent improvements on robot design understanding tasks when finetuned on our dataset than on other related ones.

References

[1] A. Gupta, S. Savarese, S. Ganguli, and L. Fei-Fei, "Embodied intelligence via learning and evolution," *Nature communications*, 2021.

[2] R. P. Ringel, Z. S. Charlick, J. Liu, B. Xia, and B. Chen, "Text2robot: Evolutionary robot design from text descriptions," in *ICRA*, 2025.

[3] G. Carbone and M. A. Laribi, *Robot Design: From Theory to Service Applications*. Springer Nature, 2022.

[4] T.-H. J. Wang, J. Zheng, P. Ma, Y. Du, B. Kim, A. Spielberg, J. Tenenbaum, C. Gan, and D. Rus, "Diffusebot: Breeding soft robots with physics-augmented generative diffusion models," *NeurIPS*, 2023.

[5] W. Li, E. Buchanan, L. K. Le Goff, E. Hart, M. F. Hale, B. Wei, M. De Carlo, M. Angus, R. Woolley, Z. Gan, *et al.*, "Evaluation of frameworks that combine evolution and learning to design robots in complex morphological spaces," *IEEE Transactions on Evolutionary Computation*, 2023.

[6] M. Li, D. Matthews, and S. Kriegman, "Reinforcement learning for freeform robot design," in *ICRA*, 2024.

[7] D. Matthews, A. Spielberg, D. Rus, S. Kriegman, and J. Bongard, "Efficient automatic design of robots," *Proceedings of the National Academy of Sciences*, 2023.

[8] W. K. Chan, P. Wang, and R. C.-H. Yeow, "Creation of novel soft robot designs using generative ai," *arXiv*, 2024.

[9] J. Hu, J. Whitman, M. Travers, and H. Choset, "Modular robot design optimization with generative adversarial networks," in *ICRA*, 2022.

[10] A. Spielberg, B. Araki, C. Sung, R. Tedrake, and D. Rus, "Functional co-optimization of articulated robots," in *ICRA*, 2017.

[11] F. Stella, C. Della Santina, and J. Hughes, "How can llms transform the robotic design process?" *Nature machine intelligence*, 2023.

[12] J. Hu, J. Whitman, and H. Choset, "Glso: grammar-guided latent space optimization for sample-efficient robot design automation," in *CoRL*, 2022.

[13] B. Song, R. Zhou, and F. Ahmed, "Multi-modal machine learning in engineering design: A review and future directions," *JCISE*, 2024.

[14] A. Heyrani Nobari, A. Srivastava, D. Gutfreund, and F. Ahmed, "Links: A dataset of a hundred million planar linkage mechanisms for data-driven kinematic design," in *IDETC-CIE*, 2022.

[15] F. Ghezelbash, A. H. Eskandari, and A. J. Bidhendi, "A dataset for mechanical mechanisms," *arXiv*, 2024.

[16] J. Xu, C. Wang, Z. Zhao, W. Liu, Y. Ma, and S. Gao, "Cad-mllm: Unifying multimodality-conditioned cad generation with mllm," *arXiv*, 2024.

[17] C. Picard, J. Schiffmann, and F. Ahmed, "Dated: Guidelines for creating synthetic datasets for engineering design applications," in *IDETC-CIE*, 2023.

[18] M. S. Khan, S. Sinha, T. U. Sheikh, D. Stricker, S. A. Ali, and M. Z. Afzal, "Text2cad: Generating sequential cad designs from beginner-to-expert level text prompts," *NeurIPS*, 2024.

[19] L. Makatura, M. Foshey, B. Wang, F. HähnLein, P. Ma, B. Deng, M. Tjandrasuwita, A. Spielberg, C. E. Owens, P. Y. Chen, *et al.*, "How can large language models help humans in design and manufacturing?" *arXiv*, 2023.

[20] D. Van Daele, N. Decleyre, H. Dubois, and W. Meert, "An automated engineering assistant: Learning parsers for technical drawings," in *AAAI*, 2021.

[21] M. T. Khan, L. Chen, Y. H. Ng, W. Feng, N. Y. J. Tan, and S. K. Moon, "Fine-tuning vision-language model for automated engineering drawing information extraction," in *ICIAI*, 2025.

[22] S. Koch, A. Matveev, Z. Jiang, F. Williams, A. Artemov, E. Burnaev, M. Alexa, D. Zorin, and D. Panozzo, "Abc: A big cad model dataset for geometric deep learning," in *CVPR*, 2019.

[23] R. Wu, C. Xiao, and C. Zheng, "Deepcad: A deep generative network for computer-aided design models," in *ICCV*, 2021.

[24] R. Wang, Y. Yuan, S. Sun, and J. Bian, "Text-to-cad generation through infusing visual feedback in large language models," *ICML*, 2025.

[25] K. D. Willis, Y. Pu, J. Luo, H. Chu, T. Du, J. G. Lambourne, A. Solar-Lezama, and W. Matusik, "Fusion 360 gallery: A dataset and environment for programmatic cad construction from human design sequences," *ACM Transactions on Graphics (TOG)*, 2021.

[26] A. Seff, Y. Ovadia, W. Zhou, and R. P. Adams, "SketchGraphs: A large-scale dataset for modeling relational geometry in computer-aided design," in *ICML Workshop*, 2020.

[27] M. T. Khan, L. Chen, Y. H. Ng, W. Feng, N. Y. J. Tan, and S. K. Moon, "Leveraging vision-language models for manufacturing feature recognition in cad designs," *JCISE*, 2025.

[28] S. Jayanti, Y. Kalyanaraman, N. Iyer, and K. Ramani, "Developing an engineering shape benchmark for cad models," *Computer-Aided Design*, 2006.

[29] H. Lee, J. Lee, H. Kim, and D. Mun, "Dataset and method for deep learning-based reconstruction of 3d cad models containing machining features for mechanical parts," *Journal of Computational Design and Engineering*, 2022.

[30] A. D. Vuong, M. N. Vu, H. Le, B. Huang, H. T. T. Binh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," in *ICRA*, 2024.

[31] X. Li, C. Mata, J. Park, K. Kahatapitiya, Y. S. Jang, J. Shang, K. Ranasinghe, R. D. Burgert, M. Cai, Y. J. Lee, *et al.*, "Llara: Supercharging robot learning data for vision-language policy," in *ICLR*, 2025.

[32] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.

[33] A. D. Vuong, M. N. Vu, B. Huang, N. Nguyen, H. Le, T. Vo, and A. Nguyen, "Language-driven grasp detection," in *CVPR*, 2024.

[34] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv*, 2023.

[35] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *NeurIPS*, 2024.

[36] M. Yavartanoo, S. Hong, R. Neshatavar, and K. M. Lee, "Text2cad: Text to 3d cad generation via technical drawings," *arXiv*, 2024.

[37] S. Redfield, "A definition for robotics as an academic discipline," *Nature Machine Intelligence*, 2019.

[38] L. Regenwetter, A. H. Nobari, and F. Ahmed, "Deep generative models in engineering design: A review," *Journal of Mechanical Design*, 2022.

[39] T. Schlagenhauf, M. Netzer, and J. Hillinger, "Text detection on technical drawings for the digitization of brown-field processes," *Procedia CIRP*, 2023.

[40] K. S. Luck, H. B. Amor, and R. Calandra, "Data-efficient co-adaptation of morphology and behaviour with deep reinforcement learning," in *CoRL*, 2019.

[41] J. Song, Y. Yang, H. Xiao, W. Peng, W. Yao, and F. Wang, "Laser: Towards diversified and generalizable robot design with large language models," in *ICLR*, 2025.

[42] Y. Jadhav and A. Barati Farimani, "Large language model agent as a mechanical designer," *Journal of Engineering Design*, 2026.

[43] Y. Zong, O. Mac Aodha, and T. M. Hospedales, "Self-supervised multimodal learning: A survey," *TPAMI*, 2024.

[44] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang, L. Guo, W. Meng, X. Zhang, *et al.*, "Multimodal fusion and vision-language models: A survey for robot vision," *Information Fusion*, 2025.

[45] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," 2021.

[46] P. Pattnayak, H. L. Patel, B. Kumar, A. Agarwal, I. Banerjee, S. Panda, and T. Kumar, "Survey of large multimodal model datasets, application categories and taxonomy," in *ICRCV*, 2025.

[47] T. Nguyen, M. N. Vu, B. Huang, A. Vuong, Q. Vuong, N. Le, T. Vo, and A. Nguyen, "Language-driven 6-dof grasp detection using negative prompt guidance," in *ECCV*, 2024.

[48] E. Kwon, F. Huang, and K. Goucher-Lambert, "Enabling multi-modal search for inspirational design stimuli using deep learning," *AI EDAM*, 2022.

[49] J. Puig Ortiz, R. Pàmies Vilà, and L. Jordi Nebot, "Exploring the application of chatgpt in mechanical engineering education," in *SEFI Annual Conference Annual Conference*, 2023.

[50] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," *arXiv*, 2022.

[51] IEEE, "IEEE Xplore Digital Library," https://ieeexplore.ieee.org/, 2025, accessed: May 12th 2025.

[52] IEEE Robotics & Automation Society, "RA-L Keywords," Software, accessed: May 12th 2025. [Online]. Available: https://www.ieee-ras.org/publications/ra-l/keywords

[53] OpenAI, "Hello GPT-4o," Software, accessed: May 12th 2025. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[54] A. Birk, "What is robotics? an interdisciplinary field is getting even more diverse," *IEEE robotics & automation magazine*, 2011.

[55] M. K. Habib and J. P. Davim, *Engineering creative design in robotics and mechatronics*, 2013.

[56] C. Clark and S. Divvala, "Pdffigures 2.0: Mining figures from research papers," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 2016.

[57] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *ICML*, 2021.

[58] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *NeurIPS*, 2019.

[59] H. Yu, Y. Tian, S. Kumar, L. Yang, and H. Wang, "The devil is in the details: A deep dive into the rabbit hole of data filtering," *arXiv*, 2023.

[60] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, 2019.

[61] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *NeurIPS*, 2024.

[62] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, *et al.*, "Siren's song in the ai ocean: a survey on hallucination in large language models," *Computational Linguistics*, 2025.

[63] Meta, "Llama 3.3 - Model Cards & Prompt formats," Software, accessed: May 12th 2025. [Online]. Available: https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

[64] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, *et al.*, "Qwen2. 5 technical report," *arXiv*, 2024.

[65] D. Khyani, B. Siddhartha, N. Niveditha, and B. Divya, "An interpretation of lemmatization and stemming in natural language processing," *Journal of University of Shanghai for Science and Technology*, 2021.

[66] X. Li, Y. Wang, and Z. Sha, "Deep learning methods of cross-modal tasks for conceptual design of product shapes: A review," *Journal of Mechanical Design*, 2023.

[67] W. Wang, K. Mrini, L. Yang, S. Kumar, Y. Tian, X. Yan, and H. Wang, "Finetuned multimodal language models are high-quality image-text data filters," *arXiv*, 2024.

[68] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *EMNLP*, 2021.

[69] T.-Y. Hsu, C. L. Giles, and T.-H. Huang, "Scicap: Generating captions for scientific figures," in *EMNLP Findings*, 2021.

[70] A. C. A. M. de Faria, F. d. C. Bastos, J. V. N. A. da Silva, V. L. Fabris, V. d. S. Uchoa, D. G. d. A. Neto, and C. F. G. d. Santos, "Visual question answering: A survey on techniques and common trends in recent literature," *arXiv*, 2023.

[71] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[72] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL Workshop*, 2005.

[73] S. Pramanick, R. Chellappa, and S. Venugopalan, "Spiqa: A dataset for multimodal question answering on scientific papers," *NeurIPS*, 2024.

[74] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *CVPR*, 2024.

[75] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," 2024.

[76] T. Li, L. Kong, X. Yang, B. Wang, and J. Xu, "Bridging modalities: A survey of cross-modal image-text retrieval," *Chinese Journal of Information Fusion*, 2024.

[77] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, 2017.

[78] S. Hartwig, D. Engel, L. Sick, H. Kniesel, T. Payer, T. Ropinski, *et al.*, "Evaluating text to image synthesis: Survey and taxonomy of image quality metrics," *arXiv*, 2024.

[79] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[80] K. M. Edwards, B. Man, and F. Ahmed, "Sketch2prototype: rapid conceptual design exploration and prototyping with generative ai," *Proceedings of the Design Society*, 2024.