CVPDL – Creative Vision-generation Preference Decision via LM

1st Huang, Wei-Hsiang
National Taiwan University
GINM

1st Zheng, Jie-Yuan
National Taiwan University
GINM

1st Su, Yu-Xuan
National Taiwan University
DS

1st Yang, Chun-Jun
National Taiwan University
GINM

***Abstract*—Evaluating the quality of images generated by advanced machine learning models is a critical challenge in computer vision and generative AI research. Traditional metrics such as Fréchet Inception Distance (FID) [1] and object detection methods often fail to capture semantic coherence, realism, and alignment with human judgment, especially in unconventional scenarios. To address these limitations, we propose a novel evaluation benchmark leveraging large language model (LLM)-based vision detection and segmentation. Our framework combines multi-modal capabilities to assess the quality of generated images across diverse object categories, including rare and culturally specific items, without reliance on predefined label sets.**

The proposed methodology integrates advanced techniques such as InstanceDiffusion [2], PixelLM [3], and Pink [4] models to enhance instance-level control and segmentation accuracy. Additionally, we establish a human judgment baseline to validate the alignment of LLM-based evaluations with perceptual realism. Experimental results suggest that our framework provides meaningful insights into image quality assessment, demonstrating correlations with human judgment. This study offers a new perspective on bridges the gap between algorithmic evaluation and human-centric AI assessment, paving the way for scalable, interpretable, and reliable benchmarks in generative AI research.

I. Introduction

Evaluating the quality of images generated by advanced machine learning models remains a critical challenge in computer vision and generative AI research. Traditional metrics such as the Fréchet Inception Distance (FID) [1] and object detectors are widely employed to benchmark generative models. However, these metrics often fail to align with human perception, particularly when assessing image coherence, realism, or semantic consistency. For example, deformed or unrealistic objects may achieve high scores due to feature similarities or high Intersection over Union (IoU) with the original bounding boxes, revealing a significant gap between quantitative metrics and qualitative human judgment.

To address these limitations, our research focuses on developing a new benchmark for evaluating generative models using large language model (LLM)-based vision detection and segmentation. By leveraging LLMs, which integrate multi-modal capabilities to process and understand both visual and textual inputs, our approach aims to provide a more human-aligned evaluation standard. This methodology enables the assessment of generative models' outputs across a diverse range of object categories, including unconventional and non-natural scenarios that are challenging for traditional metrics.

II. Related Work

The evaluation of generative models has traditionally relied on metrics such as the Fréchet Inception Distance (FID) [1] and Inception Score (IS) [5], which measure the quality and diversity of generated images by comparing their statistical properties to real images. While widely adopted, these metrics are limited in their ability to capture semantic consistency or alignment with human perception. Recent studies have highlighted their susceptibility [6] to being misled by artifacts or structural deformations that do not align with human judgment.

The advent of large language models (LLMs) and multi-modal AI systems has introduced new possibilities for image evaluation. Models such as CLIP (Contrastive Language-Image Pretraining) [7] have demonstrated the ability to align textual descriptions with visual content, enabling zero-shot evaluation tasks. Building on this foundation, recent works have integrated LLMs with vision tasks to improve contextual understanding and provide more flexible evaluation criteria.

Our work draws inspiration from these advancements, particularly in leveraging multi-modal capabilities to bridge the gap between human and machine evaluation. By incorporating LLM-based vision detection and segmentation, we extend the scope of traditional evaluation frameworks to include diverse and unconventional scenarios. Additionally, methodologies such as InstanceDiffusion [2] and PixelLM [3] have informed our approach to instance-level control and segmentation, emphasizing the importance of adaptability and scalability in modern evaluation pipelines.

III. Methodology

Our methodology aims to establish a comprehensive benchmark for evaluating generative models by integrating large language model (LLM)-based vision systems with innovative generative techniques. The proposed framework is shown in Fig 1, which contains *Scenario Prompt Generation*, *Image Generation*, *LLM-Based Vision Evaluation* and *Human Judging Baseline*. We will introduce them in details in the following.

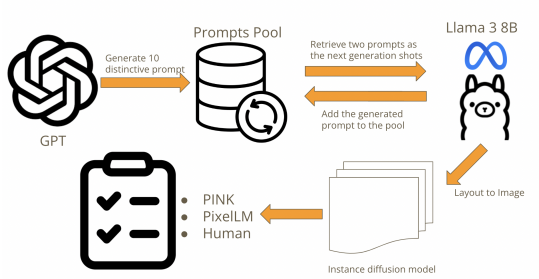

Fig. 1: The pipeline of our projects.

A. Scenario Prompt Generation

To generate various kinds of images, we leveraged the ability of pure language models to produce a total of 100 prompts with different scenaris. We first requested ChatGPT [8] to create 10 types scenario, with an overall description, and 1 to 4 annotated objects inside. The generated texts are in *json* format, shown in the next box.

Scenario JSON Format

”caption”: *Overall Scenario Description*,
”annos”: [
”caption”: *Object 1*,
”caption”: *Object 2*,
.....
]

These 10 scenarios are stored into a prompts pool. Then, we used Llama-3-8B [9] to generate the 100 target prompts via 2 shots method. The shots are randomly picked from the prompts pool and the next generated scenario will also be put into the pool. This way, we can ensure that our generated scenarios will not be restricted to only specific domains, as well as be biased by the given shots.

Finally, we will randomly assign the box positions of the annotated objects in each scenario prompt. The reason for randomly assigning the position of the objects is because we want to fully test if the final evaluations can still success no matter how strange the image is. In fact, we expect our method can be generalized to all kinds of images. An unrealistic pictures, such as the ones in some fairy tails or memes, are unavoidable. The final *JSON* format is as followed.

The generation of bounding boxes adheres to the following two conditions: For the generated 512×512 images, the size of each object is constrained such that each dimension must fall within the range of 100 to 200. The object positions are randomly assigned, with the constraint that no overlap occurs between objects.

Scenario JSON Format (Final)

”caption”: *Overall Scenario Description*,
”annos”: [
{
”caption”: *Object 1*,
”bboxes”: [xmin, ymin, width, height]
}
{
”caption”: *Object 2*,
”bboxes”: [xmin, ymin, width, height]
}
.....
]

B. Image Generation

We leveraged the ability of InstanceDiffusion model [2] to generate 8 pictures per scenario prompt, the pipeline of InstanceDiffusion is shwon in Figure 2. InstanceDiffusion introduces a novel approach to text-to-image generation by enabling precise and flexible instance-level control. It achieves this by conditioning the model on both a global caption and detailed per-instance conditions, such as location and attributes. Key components of the architecture include:

1) **Problem Definition:** The model generates images based on a global text caption c_g and per-instance conditions (c_i, l_i), where c_i describes the instance and l_i specifies its location. Locations can be expressed in various formats, including masks, bounding boxes, points, or scribbles. This generalization expands upon prior grounded text-to-image problems, offering enhanced control over scene layout and attributes.
2) **UniFusion Block:** This module bridges per-instance conditions and backbone features. It tokenizes location information using Fourier embeddings and processes text conditions through a CLIP encoder. Multiple location formats are supported, with learnable null tokens handling missing formats. The UniFusion block integrates instance embeddings with the backbone via masked self-attention, preventing information leakage between instances. Optional grounding tokens from instance masks can be added for boundary accuracy, particularly for overlapping instances or small objects.
3) **ScaleU Block:** Integrated within the UNet's decoder, ScaleU dynamically adjusts the backbone and skip-connection features using learnable, channel-wise scaling vectors. Main features are scaled directly, while low-frequency components in skip-connection features are reweighted via Fourier transformations. This improves the network's ability to respect semantic content while enhancing image quality with minimal computational overhead ($< 0.01\%$ parameter increase).
4) **Multi-instance Sampler:** To mitigate information leakage during inference, this strategy separately denoises

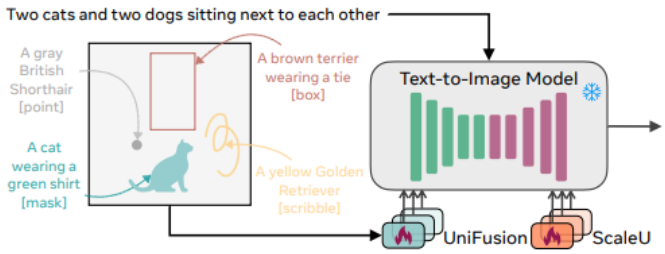

Fig. 2: The pipeline of InstanceDiffusion [2].

each instance for a fraction of the steps, generating independent latents. These latents are then averaged with global latents to create aggregated representations. The model continues denoising on this combined latent space, ensuring coherent and high-fidelity instance generation.

By utilizing InstanceDiffusion model [2] and the generated scenario prompts (Sec. III-A), we create a diverse datasets that encompass a wide range of object categories, including non-natural and unconventional scenarios resulting from the randomization of box assignments.

These datasets challenge the capabilities of generative models with the following properties:

- Complex object interactions and arrangements.
- Rare and culturally specific objects.
- Scenes with varying levels of detail and realism.

This step ensures a robust evaluation pipeline that tests the adaptability and performance of generative models in diverse contexts.

C. LLM-Based Vision Evaluation

We employ two multi-modal LLMs, PixelLM [3] and Pink [4], which we think contain the potential capablility of processing both visual and textual inputs and evaluate the quality of generated images.

- **PixelLM:** The pipeline of PixelLM is shown in Fig 3. It provided new concepts of "Seg codebook" and "Light-weight decoder". The former is a group of tokens to represent the position of an object. The predicted texts from the LLM will then be trained to produce not only texts but also the Seg codes. And the latter is used to decode these predicted codes via fusion and attention mechanism. The final output will be a mask that identified where the object is. In our project, we used the ability of the model which can output a mask from the communication. A very simple communication prompts designation is shown as follows, leading the model to output a correct mask.

Text Prompts Designation for PixelLM

Below is a situation, together with [n] objects. Please help me segment them out.
Situation: [scenario prompt here],
Objects: 1. [Object 1], 2. [Object 2], ...

We then use the predicted mask together with the input boxes to calculate the *box IoU* as well as the *prediction rate*, which is only counted if the overall size of the predicted mask is larger than 3% of the picture since our width and height are constrained in the range of 100 to 200.

- **Pink:** The pipeline of Pink is illustrated in Fig 4. It consists of three key components: the Visual Encoder, the Projection layer, and the Large Language Model (LLM). The Visual Encoder encodes an input image into a sequence of visual tokens, which are then mapped by the Projection layer into the LLM's input space. These visual tokens are concatenated with textual tokens and processed by the LLM to generate descriptive text. An adapter module was used to fine-tune both the visual encoder and the large language model simultaneously. This fine-tuning approach not only enhances the visual encoder's ability for fine-grained image understanding but also significantly reduces the number of parameters required for model fine-tuning.

Instruction Template for Pink

Please answer how many objects in the image and where, can you box it ? The objects are [Object 1], [Object 2],...

Pink evaluates the generative performance of models by combining diverse RC tasks with: **(1) Visual Encoder Integration:** Encodes input images into tokens for Referential Comprehension (RC) tasks. **(2) Self-Consistent Bootstrapping:** Ensures high-quality evaluation data by filtering noisy or inconsistent annotations in Fig 5. **(3) Task Diversity:** Covers visual relation reasoning, spatial reasoning, object counting, and object detection to comprehensively test generative capabilities.
We adapt Pink to calculate the accuracy of image generation by designing prompts to guid Pink in answering how many captions are in the image.

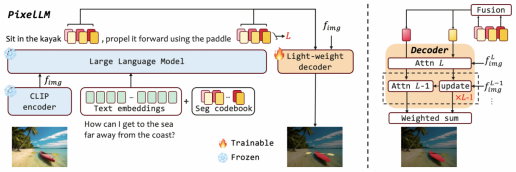

Fig. 3: The pipeline of PixelLM [3].

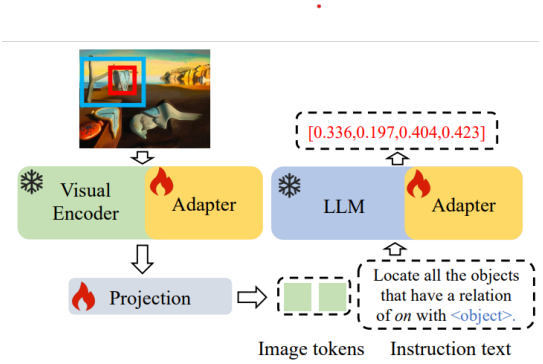

Fig. 4: The pipeline of Pink [4]

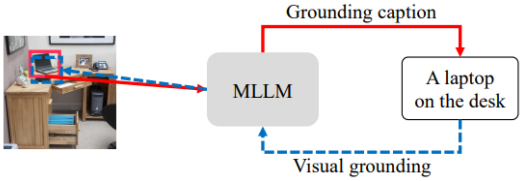

Fig. 5: The illustration of self-consistent bootstrapping method.

To summarize, PixelLM focuses on object segmentation and localization using segmentation masks, while Pink excels at combining visual understanding with language processing to perform a wide range of reasoning tasks. Together, these models provide comprehensive evaluations of image generation quality, enabling a detailed analysis of both object detection and relational reasoning capabilities.

D. Human Judging Baseline

The human evaluation criteria require that objects must be in a recognizable form as they would appear in their natural state to be considered valid. Partially incomplete but still identifiable objects are also scored. For example, a distorted human face would not be considered valid, while a generative model that produces only the lower half of a human body would still receive a score if the lower half is recognizable as such.

Figure 6 presents three examples to demonstrate the human standard. In the first image, a ”bird” and ”bench” are successfully generated in full, alongside a partially rendered ”tree.” Since the ”tree” is still recognizable, this is classified as ”all predicted.” A similar scenario occurs in the second image. However, in the third image, ”beach” and a person in the top-left corner, identifiable as ”friends,” are recognizable, while no elements related to ”selfie” are discernible. As a result, a prediction rate of ”2/3” is assigned.

IV. Contribution Attribution

We utilized *Instance Diffusion*, *PixelLM* and *Pink* as our base model. The following list the source code URL and our modification toward our projects.https://github.com/sheepjim/CVPDL_team1

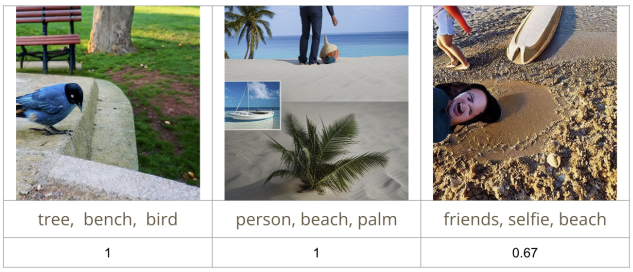

Fig. 6: Examples of human evaluation.

- **Instance Diffusion:**
  - **Source Code URL:** https://github.com/frank-xwang/InstanceDiffusion
  - **Our modification toward the projects:** We utilized the original inference code (inference.py) from the source repository with minimal modifications. Specifically, we adjusted the input JSON file to incorporate our own generated dataset, while leaving all other code and settings unchanged.
- **PixelLM:**
  - **Source Code URL:** https://github.com/MaverickRen/PixelLM
  - **Our modification toward the projects:** we modified the inference code `chat.py` to fit to the format of our generated images and extended the codes such that a continuous generation can be fulfilled. We also extracted the predicted mask labels from the model output, together with our implemented box IoU and prediction rate calculation. For postprocessing, such as comparing benchmark predictions with the human benchmark, we do not utilize any of their code.
- **Pink:**
  - **Source Code URL:** https://github.com/SY-Xuan/Pink
  - **Our modification toward the projects:** We use the inference code from source code to evaluate our images. First, we extract captions from the generated prompt JSON files and modify the input prompts by adding captions for the Pink model. The model’s output is saved to a JSON file, which contains information about the images, captions, and the number of captions in each image.

V. Experiment Results

To validate the effectiveness of our proposed benchmark, we conducted a series of experiments comparing traditional evaluation metrics with our LLM-based evaluation framework, with the result shown in Table I.

The results of PINK (0.884) and PixelLM (0.814) indicate that PINK and PixelLM are aligned with human(0.901) evaluation

standards, with their slightly lower scores reflecting a tendency to adopt even stricter criteria than humans. Next, regarding the MSE, several factors must be taken into account:

- The diversity of objects.
- The random generation of object positions and sizes, which in some cases may result in incomplete visual representation of objects, making them harder to detect.
- The majority of images contain only three objects, meaning that a single misjudgment in one criterion could lead to an MSE of approximately $0.33^2 \approx 0.1$.

Considering these challenges, we argue that the observed MSE still aligns reasonably well with human judgment standards. However, there is room for improvement in object detection models, particularly when dealing with a diverse range of objects.

Evaluation Type	Prediction Rate	Mean Square Error	Box IoU
PixelLM	0.814	0.074	0.337
Pink	0.884	0.053	xxx
Human	0.901	xxx	xxx

TABLE I: This table shows the performance of each evaluation. The prediction rate means whether the annotated objects are considered to be predicted onto the picture. The mean square error is the averaged of square of L2 distance between Human and our models. The Box IoU is the IoU between the modeled predicted masks and the original assigned boxes region.

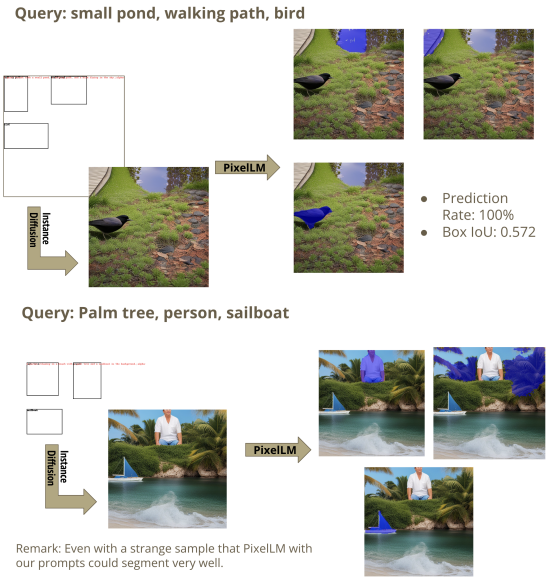

Fig. 7: The visualization example of PixelLM illustrates the capabilities of our proposed evaluation metric. The blue ones refer to the output masks. In the example above, PixelLM successfully segments objects in a natural image and provides the same evaluation results as a human observer. Even in the case of unusual images, as shown below, the predictions made by PixelLM align with human predictions. This demonstration showcases the generalization ability of this evaluation method, provided the photos are not overly unusual.

VI. Conclusion & Future Works

In this study, we proposed a novel evaluation framework to complement existing metrics in assessing the quality of images generated by advanced machine learning models. By leveraging large language models (LLMs) for vision detection and segmentation, our framework offers an alternative perspective for evaluating semantic alignment, perceptual realism, and object detection. It integrates advanced techniques, such as InstanceDiffusion, PixelLM, and Pink, enabling a comprehensive analysis of generative outputs across diverse scenarios, including rare and culturally specific objects.

Our experiments demonstrate that this framework aligns well with human judgment, providing meaningful insights into image quality assessment. The approach facilitates the evaluation of complex scenes that involve unconventional or non-natural elements, offering a more holistic view of generative model performance. While traditional metrics remain highly valuable for benchmarking, our framework contributes additional tools that enrich the evaluation process, particularly in scenarios where nuanced semantic and contextual understanding is required.

Nevertheless, there are areas where further refinement is necessary. For instance, the framework's performance in handling overlapping objects and highly abstract representations invites

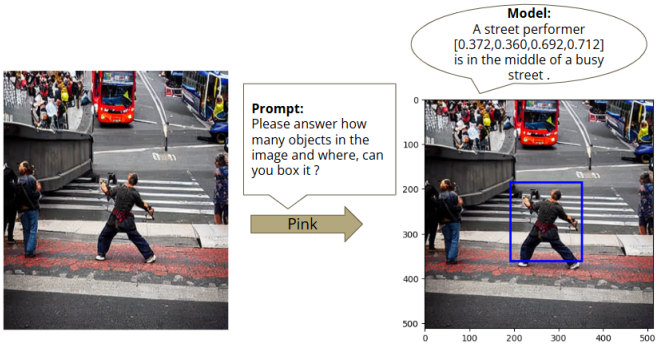

Fig. 8: An example of Pink on RC reasoning. Pink is specifically highlighted for its contribution to fine-grained image understanding and its self-consistent bootstrapping method, making it a valuable tool for detailed and accurate evaluation in multi-modal tasks.

further investigation. Moreover, as human judgment inherently varies across cultural and contextual boundaries, future work should focus on incorporating more diverse datasets and exploring methodologies that balance global and localized evaluation criteria. Expanding the framework to accommodate additional modalities, such as textual and auditory data, could also provide a more comprehensive benchmark for multi-modal generative models.

This study highlights the potential of LLM-based evaluation frameworks to complement existing approaches, bridging the gap between algorithmic evaluation and human-centric assessment. By aligning quantitative metrics more closely with human perception, we aim to contribute to the development of reliable, interpretable, and scalable benchmarks for generative AI research. As the field advances, we hope that this framework will inspire further exploration of innovative methodologies, fostering collaboration across disciplines to ensure that generative models achieve not only technical excellence but also resonance with human creativity and understanding.

VII. Team Contribution

Our project was a collaborative effort where each of the four team members contributed equally, leveraging our individual skills to achieve our shared goals. Below is an outline of each member's specific contributions and how these efforts integrated into the overall success of the project.

Member 1: Huang, Wei-Hsiang

Huang, Wei-Hsiang was in the charge of scenario prompt generation and pixelLM.

- **Scenario prompt generation**: Develop and implement the pipeline for scenario prompt generation.
- **Algorithm Implementation for PixelLM**: Designed prompts for PixelLM, manipulated it, and modified the inference code to match our need. Also, calculated the bench mark from the PixelLM prediction, including the prediction rate and the box IoU.
- **Results Analysis and Reporting**: Collaborated with team members to design and refine the user interface, including human evaluation, pipeline connection, etc.

Member 2: Zheng, Jie-Yuan

Zheng, Jie-Yuan focused on Image Generation, playing an instrumental role in the following areas:

- **InstanceDiffusion Algorithm Implementation**: Optimized the InstanceDiffusion model for generating high-quality images with precise instance-level control. Tasks included problem definition, architecture analysis, and aligning outputs with desired scene layouts and attributes.
- **Results Analysis and Reporting**: Collaborated with team members to design and refine the user interface, including human evaluation, pipeline connection, etc.
- **Active Participation in Collaboration**: Actively participated in group discussions to refine project ideas, address challenges related to evaluating generative image outputs, and align evaluation strategies with the goals of the team.

Member 3: Su, Yu-Xuan

Su, Yu-Xuan focused on *Algorithm Implementation for Pink*, playing an instrumental role in the following areas:

- **Algorithm Implementation for Pink**: Successfully implemented and optimized algorithms within the Pink framework, focusing on evaluating generative image outputs through referential comprehension tasks, such as object detection and spatial reasoning.
- **Results Analysis and Reporting**: Collaborated with team members to design and refine the user interface, including human evaluation, pipeline connection, etc.
- **Active Participation in Collaboration**: Actively participated in group discussions to refine project ideas, address challenges related to evaluating generative image outputs, and align evaluation strategies with the goals of the team.

Member 4: Yang, Chun-Chun

Yang, Chun-Chun specialized in *layout generation and Analytic tool for result estimation*, providing valuable contributions such as:

- **Algorithm Implement for layout generation**: Propose an algorithm that leverages prompts generated by LLaMA to construct corresponding layouts, format them systematically, and integrate them into an instance diffusion model.
- **Algorithm Implement for estimation**: Develop an algorithm to estimate the differences among PINK, PixelLM, and human evaluations.

VIII. Demo Video Link

Watch our demo video here:

https://youtu.be/dg-Uhv7kQs4

References

[1] 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," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6629–6640, Curran Associates Inc., 2017.

[2] X. Wang, T. Darrell, S. S. Rambhatla, R. Girdhar, and I. Misra, "Instancediffusion: Instance-level control for image generation," 2024.

[3] Y. W. Y. Z. D. F. J. F. X. J. Zhongwei Ren, Zhicheng Huang, "Pixellm: Pixel reasoning with large multimodal model," *arXiv preprint arXiv:2312.02228*, 2023.

[4] S. Xuan, Q. Guo, M. Yang, and S. Zhang, "Pink: Unveiling the power of referential comprehension for multi-modal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13838–13848, June 2024.

[5] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[6] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "Rethinking fid: Towards a better evaluation metric for image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[8] OpenAI, "Chatgpt," 2024. AI language model, accessed 2024.

[9] AI@Meta, "Llama 3 model card," 2024.