ASCIIBench: Evaluating Language-Model-Based Understanding of Visually-Oriented Text

Kerry Luo **Michael Fu** **Joshua Peguero** **Husnain Malik** **Anvay Patil**

Joyce Lin **Megan Van Overborg** **Ryan Sarmiento** **Kevin Zhu**

Algoverse AI Research
kerryluo1@gmail.com, kevin@algoverseairesearch.org

Abstract

Large language models (LLMs) have demonstrated several emergent behaviors with scale, including reasoning and fluency in long-form text generation. However, they continue to struggle with tasks requiring precise spatial and positional reasoning. ASCII art, a symbolic medium where characters encode structure and form, provides a unique probe of this limitation. We introduce ASCIIBench, a novel benchmark for evaluating both the generation and classification of ASCII-text images. ASCIIBench consists of a filtered dataset of 5,315 class-labeled ASCII images and is, to our knowledge, the first publicly available benchmark of its kind. Alongside the dataset, we release weights for a fine-tuned CLIP model adapted to capture ASCII structure, enabling the evaluation of LLM-generated ASCII art. Our analysis shows that cosine similarity over CLIP embeddings fails to separate most ASCII categories, yielding chance-level performance even for low-variance classes. In contrast, classes with high internal mean similarity exhibit clear discriminability, revealing that the bottleneck lies in representation rather than generational variance. These findings position ASCII art as a stress test for multimodal representations and motivate the development of new embedding methods or evaluation metrics tailored to symbolic visual modalities. All resources are available at `https://github.com/ASCIIBench/ASCIIBench`.

1 Introduction

Scaling language models has been shown to induce emergent capabilities [Wei et al., 2022], including those involving positional understanding, such as the generation and editing of `TikZ` drawings [Bubeck et al., 2023]. We define ASCII art as the intersection of text and vision. The generation and classification of ASCII art introduces challenges that are distinct from conventional NLP and multimodal benchmarks: characters function as visual primitives rather than semantic tokens, necessitating strict structural regularity seen in other forms of structured data like tables [Chen, 2022]. In contrast to natural images, ASCII art is both present in the pretraining distribution of unimodal language models and natively aligned with their tokenization schemes, enabling direct evaluation without additional adaptation.

2 The ASCIIBench Dataset

We introduce ASCIIBench, a high-quality benchmark for ASCII art understanding and generation. Sourced ethically from `ascii.co.uk`, the data underwent a rigorous multi-stage curation pipeline.

Preprint.

The final dataset contains 5,315 unique ASCII art pieces across 752 classes (e.g., `aircraft`, `birds`). All art is credited to the original creators on `ascii.co.uk`. In the absence of explicit licensing, we adhered to standard research practices described in Appendix A.

2.1 Data Curation & Analysis

Raw ASCII art contains pervasive noise like signatures and tags. An in-depth description of our data cleaning methodology and dataset analysis can be found in Appendix B and Appendix C.

3 Classification

Models We evaluated multiple models on classification and generation tasks, including Llama 3-8B, Llama 3-8B-Instruct, GPT-3.5, GPT-4o, GPT-4o-mini, GPT-5-mini, and Claude 3.5 Sonnet, testing text-only, vision-only, and text-vision prompts to compare performance across modalities.

3.1 Model Testing Procedure

```
Classify the following ASCII art into one of the given categories:

        _.---._    /\\
     ./'       "--`\//
   ./              o \
  /./\  )______   \__ \
 ./  / /\ \   | \ \  \ \
    / /  \ \  | |\ \  \7
     "    "    "  "

Choices: puppy, aardvark, giraffe, tapir
Classification: giraffe
Predicted_class: giraffe, Actual_class: aardvark, Correct: false
```

Figure 1: Example classification prompt with result

Prompt ASCII images are preprocessed based on input modalities. Image preprocessing is described in Appendix F. The model is then prompted to select one of four choices in the format shown in Figure 1.

Evaluation Metrics Performance was measured by the model's macro and micro accuracy

4 Generation

Models We prompted GPT-3.5, GPT-4, and GPT-4o to generate 5 ASCII images for each class.

Approach To evaluate fidelity, we require an image-to-image metric that captures both the visual and textual characteristics of ASCII art. We use CLIP [Radford et al., 2021], which aligns images and text through large-scale contrastive training. By comparing embeddings of generated images to reference embeddings derived from ground-truth data, we assess generation accuracy.

4.1 Evaluation Metrics

We leverage CLIP cosine similarity between generated and reference images and representation quality of the embedding space using alignment and uniformity [Wang and Isola, 2022]. We report ROC-AUC for same-class retrieval in Section 5.2. ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) quantifies how well a model separates positive from negative pairs, with 0.5 indicating random performance and 1.0 indicating perfect discrimination.

4.2 CLIP Cosine Similarity

Purpose CLIP cosine similarity is a metric used to evaluate how similar two images are in the context of their high-level features extracted by the CLIP model.

Implementation ASCII art is rendered following the steps in Appendix F and then embedded with CLIP. The CLIP model, known for its ability to understand high-level visual concepts through natural language supervision, is used to process these images [Radford et al., 2021] . The model extracts feature vectors representing the semantic content of each image. The cosine similarity score ranges from -1 (completely different) to 1 (exactly the same), with higher scores indicating greater similarity [Radford et al., 2021].

4.3 Alignment & Uniformity

Alignment measures intra-class compactness, while uniformity quantifies dispersion in the embedding space [Wang and Isola, 2022]. Out-of-the-box CLIP shows alignment of 5.85 (squared 34.20). Fine-tuning increases alignment to 8.90 (squared 79.16) and improves uniformity from baseline to -7.61 (t=1), -8.09 (t=5), and -8.21 (t=10). Together with stable cosine similarities, these results confirm that CLIP is not experiencing representation collapse.

5 Results

5.1 Classification Results

We evaluate the performance of various models when classifying ASCII art using the methods in Section 3.1 with a maximum of 50 output tokens. We report results across three modalities: T (text-only), V (vision-only), and T+V (text+vision). Responses were filtered for possible string parsing errors, resulting in a <2% average removal. Unfiltered and filtered results are shown in Table 1.

Table 1: Model performance comparison on raw (left) and filtered (right) datasets.

Raw (Unfiltered) Dataset					**Filtered Dataset**				
Model	Mod.	Micro acc. (%)	Macro acc. (%)	Pass rate (%)	Model	Mod.	Micro acc. (%)	Macro acc. (%)	Pass rate (%)
LLaMA3.1-8B-Inst	T	34.27	31.89	91.78	LLaMA3.1-8B-Inst	T	34.50	32.01	91.69
LLaMA3.1-8B	T	29.00	25.07	82.67	LLaMA3.1-8B	T	29.39	25.40	82.69
GPT-5-mini	T	61.60	62.39	**99.38**	GPT-5-mini	T	61.36	61.97	**99.35**
	V	77.25	**84.13**	99.24		V	77.25	**84.13**	99.24
	T+V	73.27	73.84	99.01		T+V	73.27	73.84	99.01
GPT-4o-mini	T	73.61	77.60	95.23	GPT-4o-mini	T	73.52	77.27	95.19
	V	75.72	77.77	97.27		V	75.72	77.77	97.27
	T+V	76.02	77.55	96.67		T+V	76.02	77.55	96.67
GPT-4o	T	75.44	80.23	96.63	GPT-4o	T	75.64	80.26	96.61
	V	**77.49**	82.16	98.75		V	**77.49**	82.16	98.75
	T+V	76.56	79.74	98.52		T+V	76.02	77.55	96.67
GPT-3.5-turbo	T	39.05	33.54	91.34	GPT-3.5-turbo	T	39.98	33.77	91.31
Claude-3.5-Sonnet	T	59.55	56.98	98.54	Claude-3.5-Sonnet	T	59.84	57.23	98.65
	V	76.40	76.92	99.08		V	76.40	76.92	99.08
	T+V	76.48	76.89	99.08		T+V	76.48	76.89	99.08

5.1.1 Interpretation

Our results align with those of Jia et al. [2024]. Larger models had greater performance, and all accuracy values were over 25%, indicating that models did not choose arbitrarily. Across both raw and filtered datasets, we find that vision-only models consistently outperform text-only and text+vision counterparts, with GPT-4o achieving the highest macro accuracy at 82.2%. Text-only performance lags significantly, especially for LLaMA and GPT-3.5, underscoring the difficulty of modeling ASCII art as pure text. Surprisingly, adding text to vision does not improve performance and in some cases degrades it, suggesting that current multimodal fusion strategies do not capture ASCII structure effectively. Filtering has little effect on overall trends, indicating robustness of the observed modality gaps.

5.2 Generation Results

On unfiltered generations, CLIP showed weak class separation (ROC-AUC ≈ 0.55; silhouette -0.46), and t-SNE revealed no clear clusters. After filtering inconsistent generations (std > 0.15, mean similarity < 0.3), ROC-AUC rose to 0.83, demonstrating that CLIP can discriminate effectively when ASCII generations are semantically consistent. This indicates that the bottleneck lies in the quality of LLM-generated ASCII rather than in the evaluator.

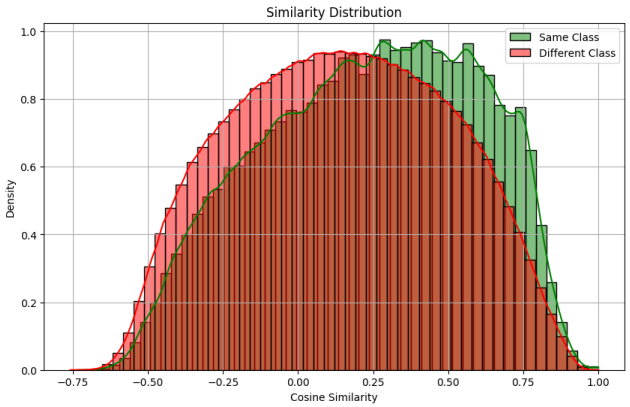

Figure 2: Cosine similarity distributions. Green indicates positive (intra-class) pairs, red indicates negative (inter-class) pairs.

Shown in Figure 2, while there is a separation between inter- and intra-class distributions, the separation is not complete, with substantial overlap remaining.

5.2.1 CLIP Representation Analysis

We examined cosine similarities, silhouette scores, and t-SNE visualizations of CLIP embeddings (Figure 5). Out-of-the-box CLIP produced weak intra- and inter-class separation (AUC ≈ 0.558, silhouette -0.46), and fine-tuning with triplet loss yielded only modest gains. Filtering noisy generations did not resolve this, as even the lowest-variance subsets approached near chance performance (AUC = 0.641). However, when restricting analysis to classes with high mean similarity, AUC increased to 0.83, indicating that CLIP can represent ASCII structure only for a subset of well-formed categories. These results show that the primary limitation lies in CLIP's representational capacity for ASCII art, along with variance in model generations.

6 Limitations

Our findings show that evaluation quality depends strongly on input consistency. CLIP performs well only when generations are visually coherent and semantically aligned, but typical LLM outputs are noisy and inconsistent, especially for vague categories. This highlights a dual bottleneck: the instability of ASCII generation and the limited ability of a broad, general-purpose model like CLIP to represent ASCII structure. Filtering demonstrates an upper bound of performance but is not a sustainable evaluation strategy, as it amounts to testing on inputs already close to the training distribution. Future work should explore specialized, smaller models, which may capture ASCII-specific patterns more effectively than CLIP.

7 Conclusion

We introduce ASCIIBench, a benchmark for evaluating ASCII art on classification and generation tasks, and used it to probe how multimodal models represent symbolic visual inputs. Empirically, vision-only models consistently outperform text-only and text+vision settings on classification, while CLIP-based evaluation of generations provides limited class separation on unfiltered outputs and improves primarily for classes with high internal similarity. These trends position ASCII art as

a stringent stress test for multimodal reasoning: performance hinges on both the consistency of generations and the representational suitability of the embedding model for ASCII structure. Looking ahead, we advocate standardized rendering and preprocessing protocols to enable fair cross-model comparisons, improved prompting and training strategies for ASCII generation, and exploration of structure and variance-aware metrics to better capture and evaluate symbolic layout.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712, 2023. URL `https://api.semanticscholar.org/CorpusID:257663729`.

Wenhu Chen. Large language models are few(1)-shot table reasoners. *ArXiv*, abs/2210.06710, 2022. URL `https://api.semanticscholar.org/CorpusID:252872943`.

Moonjun Chung and Taesoo Kwon. Fast text placement scheme for ascii art synthesis. *IEEE Access*, 10:40677–40686, 2022. doi: 10.1109/ACCESS.2022.3167567.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Qi Jia, Xiang Yue, Shanshan Huang, Ziheng Qin, Yizhu Liu, Bill Yuchen Lin, and Yang You. Visual perception in text strings, 2024. URL `https://arxiv.org/abs/2410.01733`.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms, 2024.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review, 2018.

Rémi Kazmierczak, Gianni Franchi, Nacim Belkhir, Antoine Manzanera, and David Filliat. A study of deep perceptual metrics for image quality assessment, 2022.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

Kazuyuki Matsumoto, Akira Fujisawa, Minoru Yoshida, and Kenji Kita. Ascii art classification based on deep neural networks using image feature of characters. *J. Softw.*, 13:559–572, 2018. URL `https://api.semanticscholar.org/CorpusID:53281518`.

Katsunori Miyake, Henry Johan, and Tomoyuki Nishita. An interactive system for structure-based ascii art creation. 01 2011.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020.

Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale

Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

Xuemiao Xu, Linling Zhang, and Tien-Tsin Wong. Structure-based ascii art. In *ACM SIGGRAPH 2010 Papers*, SIGGRAPH '10, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450302104. doi: 10.1145/1833349.1778789. URL https://doi.org/10.1145/1833349.1778789.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks, 2023.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. doi: 10.1109/JPROC.2020.3004555.

Appendix

A Data Sourcing

We express our gratitude to the ASCII artists. We made slight modifications to the original ASCII art and provide the URL to the source of our data. Our dataset is licensed under CC BY NC 4.0.

B Data Curation

We developed a custom web crawler and an 11-step automated pipeline to remove these artifacts, followed by a multi-stage manual review described in Appendix E. Abstract or ambiguous categories

(e.g. "small") were excluded, retaining only well-defined classes. Three annotators then applied a strict rubric to eliminate pieces with: (1) inappropriate content, (2) excessive intra-class variation, (3) overly complex structures, or (4) low quality. This conservative process, requiring strong annotator agreement, removed over 13,000 low-quality images and 1,800 ambiguous classes, resulting in a focused, high-quality benchmark.

C Data Analysis

The curated dataset exhibits a natural long-tail class distribution (Figure 3). The largest categories are aircraft (13.3%), land transportation (11.1%), and birds (10.4%) (Figure 4). A t-SNE visualization of class embeddings (Figure 5) confirms semantic coherence, showing clear clustering of related concepts (e.g. animals), demonstrating that ASCII art encodes learnable semantic structures. Character frequency analysis (Figure 7) reveals the artistic "vocabulary": the space character is dominant (>1.6M occurrences), followed by structural elements like `-`, `|`, and `_`. Alphanumeric characters are used sparingly as accents.

D Dataset Analysis Figures

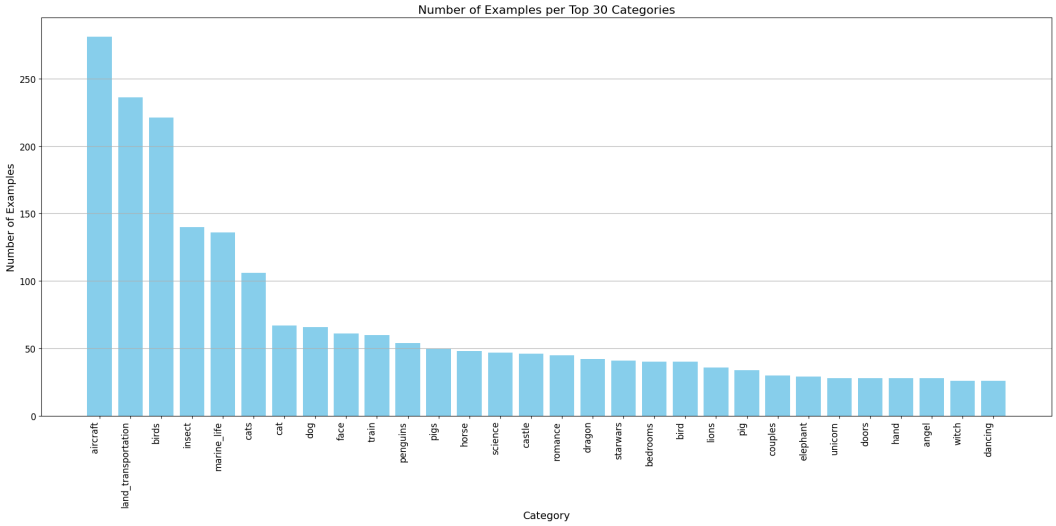

Figure 3: Top 30 Class Histogram

E Noise Removal Pipeline

1. Tags consisting of three or fewer alphabetic characters are replaced with whitespace.
2. Creator tags in the last three lines of the artwork are detected and removed while preserving structural spacing.
3. Non-visible Unicode control characters are filtered out.
4. Alphanumeric creator tags appended to the end of the ASCII art are removed.
5. Full names and abbreviated creator tags (e.g., `'Matthew Kenner'` or `'jps'`) positioned on the right margin are eliminated.
6. Tags labeled as `'unknown'` are discarded.
7. Left-aligned creator signatures are identified and removed.
8. Common date formats (e.g., `12/21/2023`, `21nov2023`, `12.21.2023`) are detected via expression matching and subsequently stripped.
9. Contact information, such as email addresses, is localized and pruned.

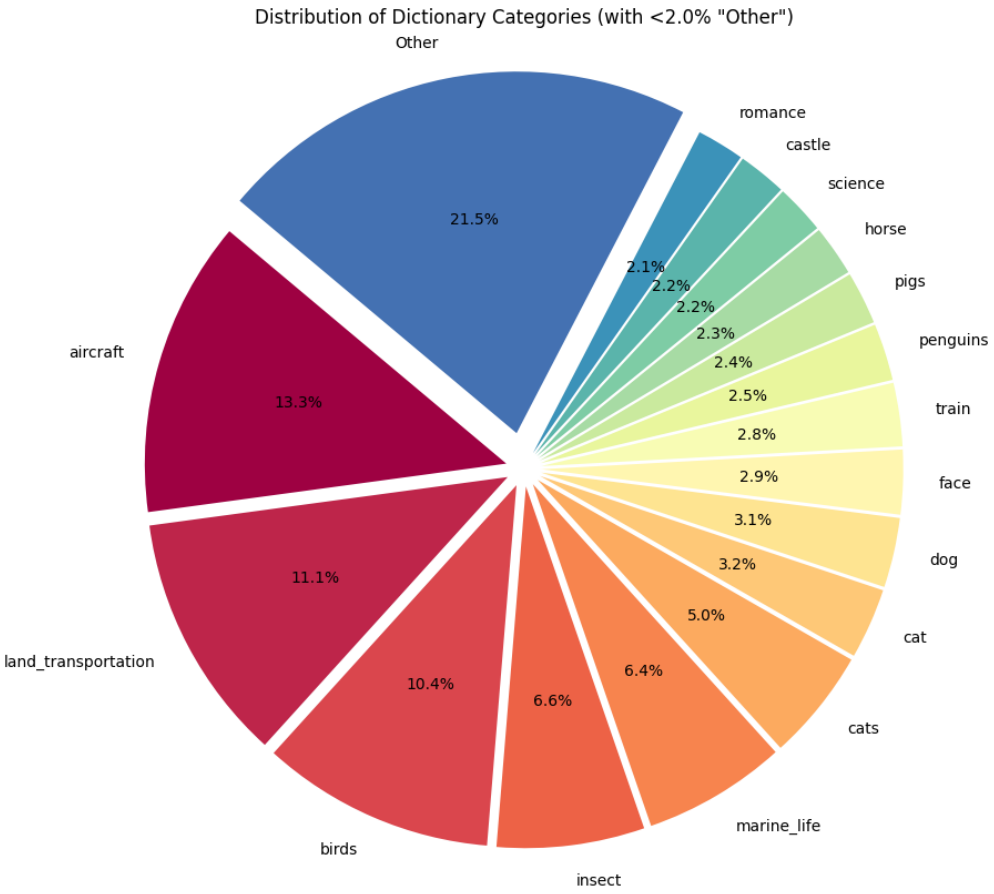

Figure 4: Class Distribution Pie Chart

10. Three-letter creator tags enclosed in dashes (-) or brackets ([]) are filtered out.
11. Known problematic signatures, maintained in a blacklist, are systematically removed.

F Preprocessing

We follow Jia et al. [2024], using a black monospaced font (DejaVu Sans Mono) on a white background. No blur is added to preserve structural integrity.

G Non-Monospaced Font Ablation

To examine whether models rely on spatial alignment rather than textual content, we replaced the fixed-width (monospaced) ASCII font with a proportional (non-monospaced) one. As shown in Table 2, the results remain nearly unchanged, suggesting that models primarily depend on optical character recognition (OCR)-like mechanisms rather than explicitly reasoning about the positional structure of characters.

H Related Works

Emergent Behaviors: LLMs have been studied for their emergent properties as they scale. Wei et al. [2022] highlights that larger models, ones with more parameters and diverse data, possess emergent abilities such as improved reasoning and fluency in text generation. Our work extends this

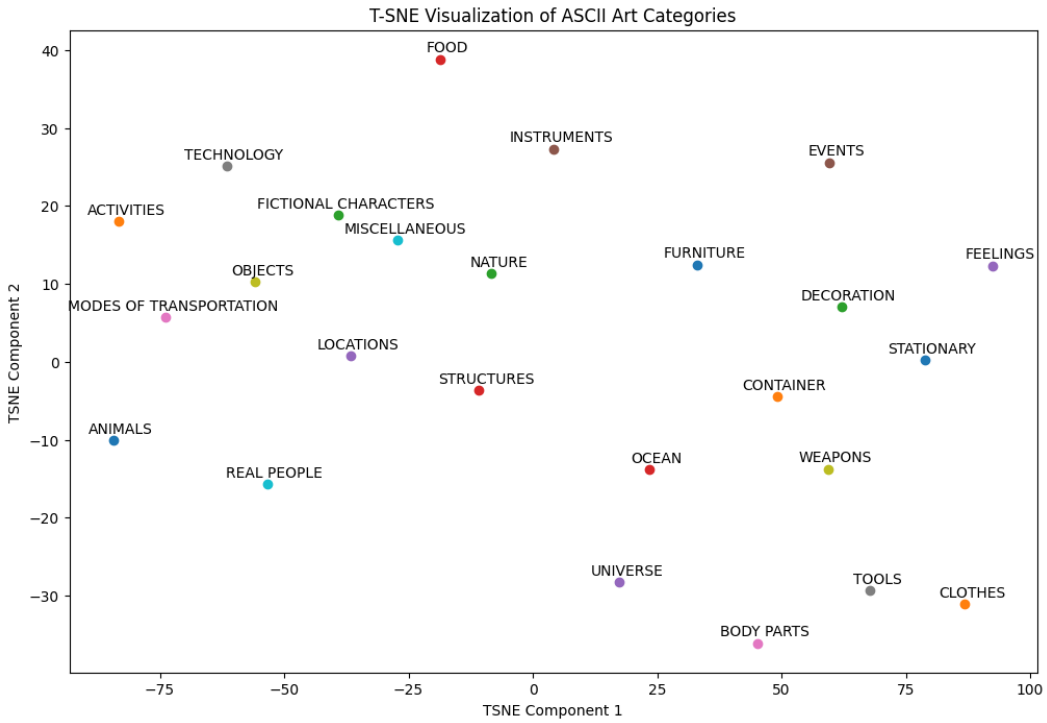

Figure 5: T-SNE visualization of class embeddings

Model	Modality	Accuracy
GPT-5	Vision & Text	0.7057
GPT-5	Vision Only	0.7118

Table 2: Performance on ASCII art rendered in non-monospaced font. The minimal difference suggests reliance on OCR-like recognition rather than positional reasoning.

to ASCII-text image generation, which requires textual understanding and visual creativity, skills not typically emphasized in standard LLM evaluations.

LLM ASCII Word Recognition: The Beyond the Imitation Game benchmark (BIG-bench) introduced by Srivastava et al. [2023] addresses the need to understand the capabilities of LLMs across various tasks. Our research focused on its ASCII word recognition dataset. Similarly, the ArtPrompt jailbreak attack highlights the need for the improvement of LLM performance in identifying ASCII text, an ability crucial to prevent the circumvention of safeguards and elicitation of unintended behaviors [Jiang et al., 2024].

Other Benchmarks: Recent efforts by Jia et al. [2024] have explored the visual perception capabilities of large language and multimodal models through ASCII art, introducing the *ASCIIEval* benchmark and a corresponding training set to evaluate models' ability to interpret visual semantics in text strings. Their study establishes ASCII art as a modality-agnostic medium bridging textual and visual understanding, offering a valuable testbed for analyzing perception and modality fusion in modern models. However, while Jia et al. [2024] constructed the dataset and presented detailed benchmarking results, the accompanying resources were not publicly released at the time of publication. In contrast, our work provides the first publicly available implementation and dataset release for ASCII-based multimodal reasoning, enabling full reproducibility and community-driven extensions of this emerging research direction.

Vision-Language Integration and Multimodal Models: GPT-4V(ision) produces human-aligned scores with detailed explanations, showing promise as a universal automatic evaluator despite some limitations [Zhang et al., 2023]. Flamingo models demonstrate strong few-shot learning capabilities,

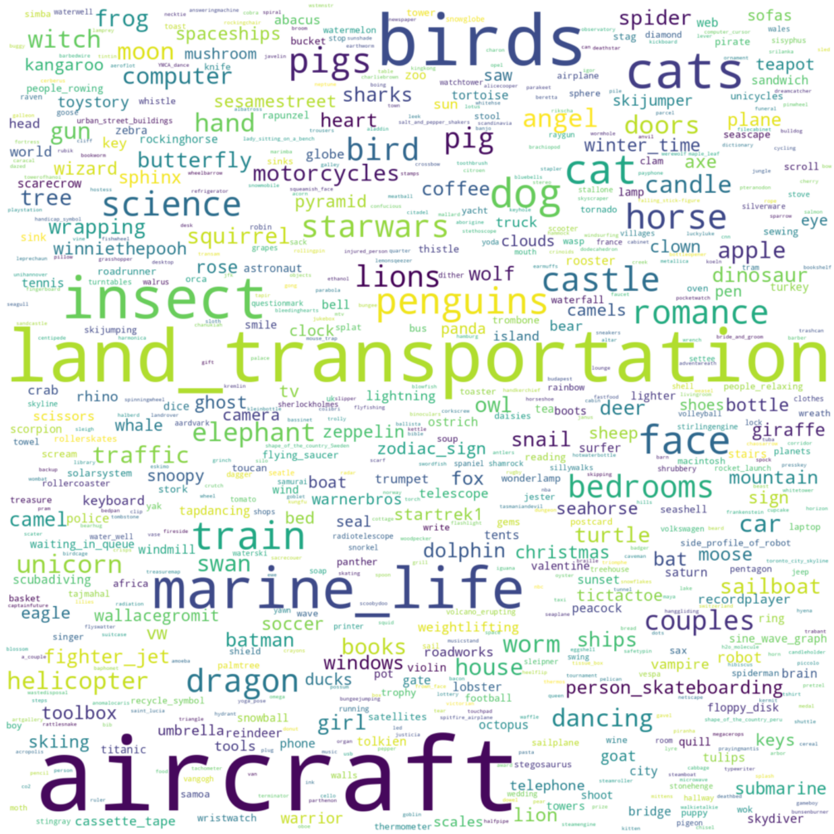

Figure 6: Word cloud of class labels

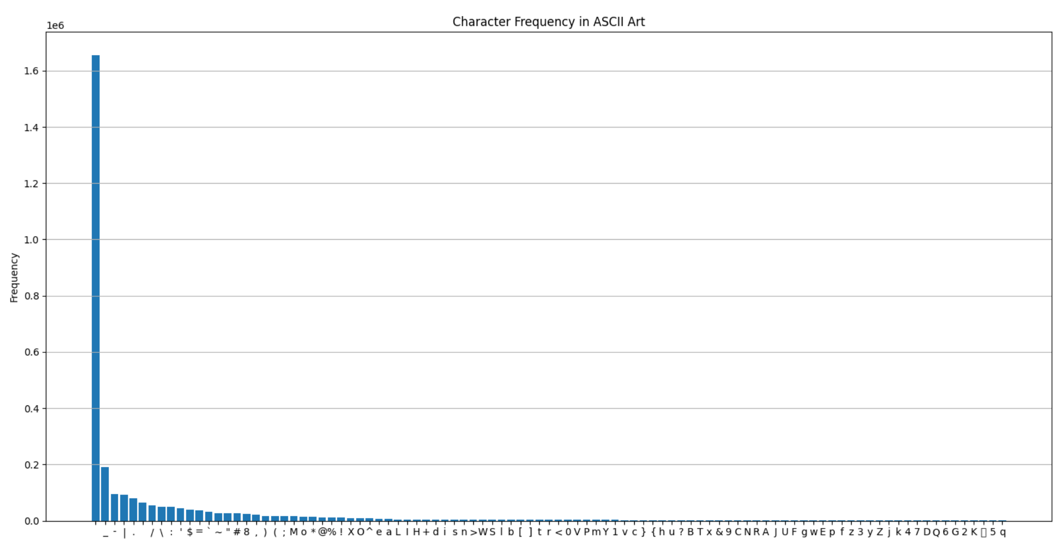

Figure 7: Character Frequency Histogram

showcasing potential to give LLMs adaptive abilities and decreased dependence on large task-specific datasets [Alayrac et al., 2022]. Ramesh et al. [2021] introduces an autoregressive transformer-based approach for text-to-image generation with competitive zero-shot performance compared to domain-specific models. These capabilities lead into ASCII art generation, which poses unique

challenges due to merging textual and visual information. However, the advanced multimodal reasoning of these models also introduces new vulnerabilities.[Jia et al., 2024] demonstrate that Large Vision-Language Models (LVLMs) like LLaVA and GPT-4V are highly susceptible to *Self-Generated Typographic Attacks*, where the model itself is leveraged to create deceptive text and descriptions that cause misclassification, reducing accuracy by up to 60%. This underscores a critical weakness in how LVLMs fuse and weight visual and textual signals. Prior works have explored the generation of stylized textual outputs using neural networks, with some focusing specifically on the artistic transformations of text and image data [Matsumoto et al., 2018]. Models like ViLBERT and VisualBERT use multimodal pre-training to boost performance in vision-language tasks [Li et al., 2019] [Lu et al., 2019]. They guide our fine-tuning of the CLIP model for effective ASCII art generation from text descriptions.

Evaluation Metrics and Methodologies: Finally, our evaluation of LLM outputs uses standard metrics used in both language and image processing domains. Metrics such as FID scores, typically used to assess image quality, were adapted to assess the uniqueness and clarity of ASCII-text images produced by our model. We employ CLIP for its ability to effectively bridge text and image representations. Research by Radford et al. [2021] demonstrates its robust performance in zero-shot classification tasks. This makes CLIP ideal for our needs, as our models must both generate and classify ASCII images from minimal prompts.

H.1 Other ASCII generation methods

The exploration of AI in the context of ASCII art has witnessed growing interest in recent years, with researchers exploring methods to optimize conversion accuracy. There are many notable contributions in this field. [Goodfellow et al., 2014] proposed a new framework for estimating generative models via an adversarial process. Researchers have delved into the use of GANs to generate realistic and well-designed ASCII art. By leveraging the adversarial training paradigm, these models can produce a large range of high-quality ASCII representations. Similarly, [Gatys et al., 2015] explored Convolutional Neural Networks and their ability to create artistic imagery by experimenting with the style and content of an image, also known as Neural Style Transfer. [Jing et al., 2018] provides an extension of this idea, comparing different Neural Style Transfer qualitatively and quantitatively. They discuss applications of NST and problems to be addressed in future research. Studies have investigated how deep neural networks can be trained to transfer artistic styles onto ASCII images, showcasing the potential for creative synthesis.

H.1.1 Transfer Learning

Naturally, it is also extremely important for the models to efficiently produce accurate results. [Zhuang et al., 2021] reviews more than forty representative transfer learning approaches from a data and model perspective. Their paper provides more than twenty experiments of different learning models' performances. The exploration of the efficacy of transfer learning in training models for ASCII art generation and leveraging pre-trained models on large datasets allow researchers to enhance the efficiency and artistic quality of AI-generated ASCII images.

H.1.2 Methods & Metrics

Our research paper took inspiration from methods mentioned in [Pang et al., 2021], which explored developments in I2I translations and analyzed key techniques to elaborate on the effect of I2I on the research and industry community. The paper introduced the problem setting of the image-to-image translation task, introduced the generative models used for I2I methods, and discussed the work and applications of multi-domain I2I tasks. Many of these methods and metrics mentioned in this paper are parallel to evaluation metrics we implemented, but while this paper mainly evaluated methods on I2I, our research drew comparisons between these methods and ASCII image generation standards.

H.1.3 Frechet Inception Distance

More specifically, Frechet Inception Distance, which produces an FID score, has been the most widely used metric for measuring the similarity between real and generated images. Muhammad [Naeem et al., 2020] focuses on the reliability of certain methods. They concluded that while variants of precision and recall metrics are generally unreliable methods, density and coverage metrics will

provide more interpretable and reliable comparisons. While precision metrics can overestimate the manifold around real outliers, density fixes this issue by more accurately representing the distribution around real samples. The objective of coverage is to improve upon the recall metric. When models generate many unrealistic and diverse samples, this can skew the data and lead to a false increase in the recall measure. Coverage addresses this by building the manifolds around real samples as opposed to fake ones. This approach is less prone to overestimation since real samples tend to have fewer outliers compared to generated samples. The goal for our purposes would be to capture how well the generated ASCII art (fake samples) represents the original ASCII art (real samples) in both details (density) and overall composition (coverage).

H.1.4 Interactive System for Structure-based ASCII Art Creation

Interactive structure-based systems [Xu et al., 2010] invite users to actively participate in the creation of ASCII art; the interactive paradigm empowers users to foster a collaborative synergy between human creativity and computational assistance. While it is an earlier paper, [Miyake et al., 2011] proposes to input images divided into grids for glyph matching using four metrics employed for converting images into ASCII art: template matching which considers pixel positions for dissimilarity measure, normalized cross-correlation which minimizes the influence of line width differences using histograms, Histogram of Oriented Gradients (HOG) representing line directions, and distance transformation indicating line positions. In comparison, our work focuses mainly on AI generation, so while it is not as collaborative it focuses on the optimization and comparison of methods for comparing the output of ASCII art utilizing either tone-based style, which is a detailed and comprehensive image, or structure-based style, which is a simple outline of the image using less characters.

H.1.5 Perceptual Metrics for Image Quality Assessment

The recent success of perceptual messages based on deep neural networks in regards to the Image Quality Assessment (IQA) task [Kazmierczak et al., 2022] has led to a growing interest in new metrics that outperform previous metrics to develop perceptual information at different resolutions. Whereas the IQA metric is generally easily perceivable for humans, it is more difficult to set a metric for a computational algorithm. Our work investigates the model's abilities to generate accurate images by comparing it to Euclidean distance and the SSIM index, groundwork laid by [Chung and Kwon, 2022].

H.1.6 ASCII Representation Learning

While prior work has explored ASCII conversion from images [Matsumoto et al., 2018], little attention has been given to understanding how models internally represent ASCII structures. Our probing of CLIP embeddings extends this line of inquiry, revealing architecture and how ASCII images are represented in the model.