Draw Me Like My Triples: Leveraging Generative AI for Wikidata Image Completion (Poster)

Raia Abu Ahmad, Martin Critelli, Şefika Efeoğlu, Eleonora Mancini, Célian Ringwald, Xinyue Zhang, Albert Meroño-Peñuela

Draw Me Like My Triples: Leveraging Generative AI for Wikidata Image Completion

Raia Abu Ahmad[1], Martin Critelli[2], Şefika Efeoğlu[3,4], Eleonora Mancini[5], Célian Ringwald[6], Xinyue Zhang[7], Albert Meroño-Peñuela[8]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH), Berlin, [2]University Ca'Foscari of Venice, [3]Freie Universität Berlin, [4]Technische Universität Berlin, [5]University of Bologna [6]Université Côte d'Azur, Inria, CNRS, I3S [7]University of Oxford, Oxford [8]King's College London

Abstract

We leverage generative AI for the task of creating images for Wikidata items that do not have them. Our approach uses knowledge contained in Wikidata triples of items describing fictional characters and uses the fine-tuned T5 model based on the WDV dataset to generate natural text descriptions of items about fictional characters with missing images. We use those natural text descriptions as prompts for a transformer-based text-to-image model, Stable Diffusion (SD) v2.1, to generate plausible candidate images for Wikidata image completion. We motivate this choice by the fact that querying Wikidata shows that only **7%** out of the 83.7K instances of the fictional character class have an image.

Our work addresses the following **Research Questions (RQs)**:

- **RQ1**: To what extent can different types of prompts based on triples be used in text-to-image models to produce high-quality images?
- **RQ2**: To what extent can the output of generative AI be used for Wikidata image completion?
- **RQ3**: How can generative text-to-image models be evaluated?

Proposed Approach

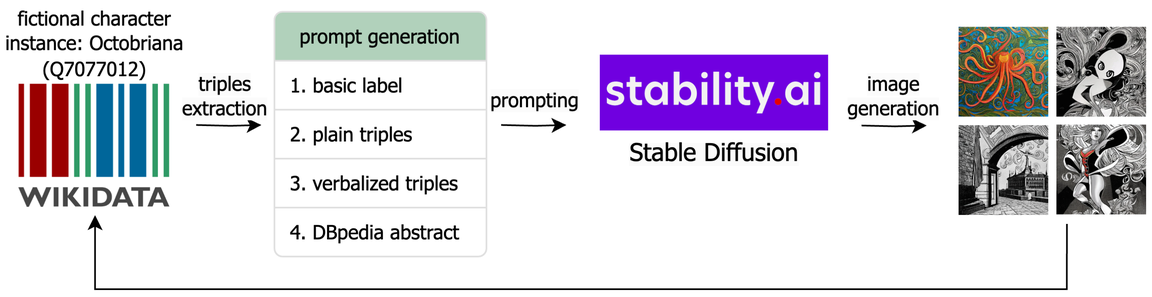

Figure 1: The pipeline of our proposed KG completion process.

1. Triple extraction

- Generating an image for a specific character requires a description gathered from its related triples.
- Obtained through SPARQL queries in Wikidata, yielding all triples with the character.

2. Prompt Generation

- **Basic Label**: *Label* of the fictional character on Wikidata.
- **Plain Triples**: derived by concatenating the subject, predicate, and object of a triple to form a single sentence, utilising all available triples linked to a specific entity. Notably, sentences might lack proper structure and grammar.
- **Verbalised Triples**: transformation of structured data (i.e., triples) into human-readable formats (i.e., text) which obtained them by using the T5 language model fine-tuned on the WDV dataset [1].
- **DBpedia Abstracts**: obtained by querying the English chapter of DBpedia [2] which is the only one originally written by a human in natural language.

3. Images Generation

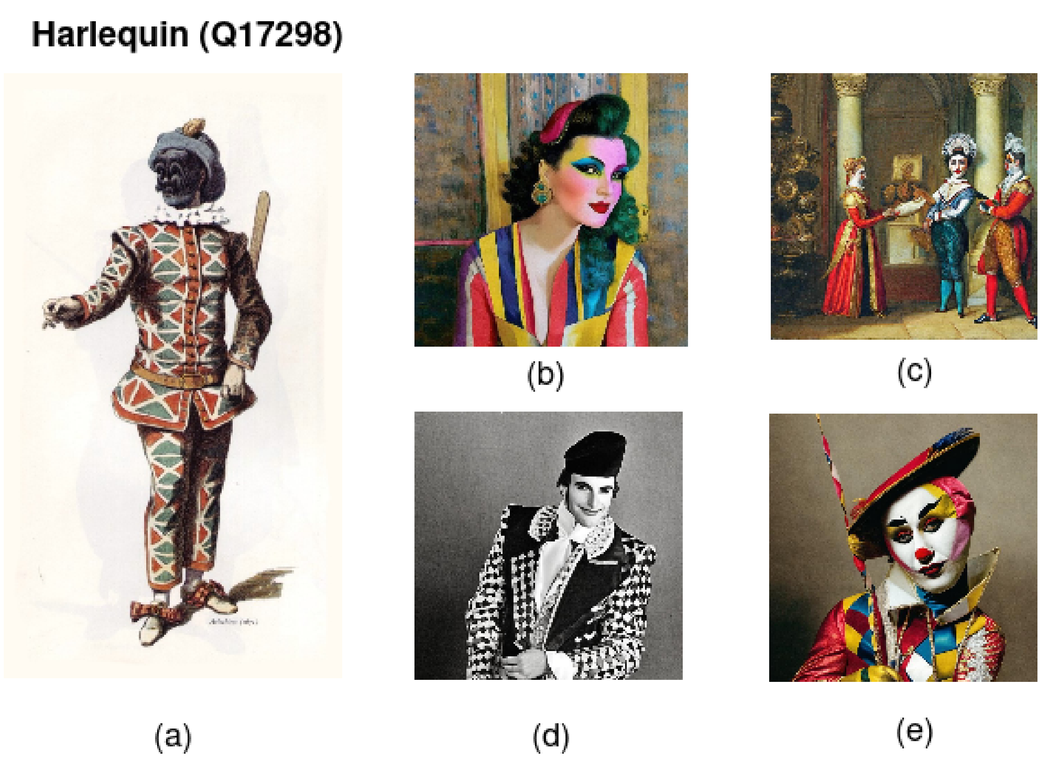

Figure 2: Images for the character of Harlequin. (a) Ground truth from Wikidata. (b) the basic label prompt. (c) the plain triples prompt. (d) verbalised triples prompt. (e) DBpedia abstract prompt

- Utilizing SD version 2.1; however, limited to the English language.
- The SD model requires negative prompts to eliminate its malformation issues, so integrated negative prompts proposed on a public practitioners community GitHub repository: https://github.com/mikhail-bot/stable-diffusion-negative-prompts
- Resolved the limitation on the number of tokens in the prompt sentences by Compel library: https://github.com/damian0815/compel

Evaluation

Automatic evaluation

- Compared the generated images with the ground-truth one by computing the two following metrics:
  1. UQI [7] is based on pixel-based
  2. CLIPscore [4] exploits the joined text-to-image CLIP embedding used by the SD model.
- The CLIPscore records a significant difference between the prompts used for the generation.

Prompt Type	min ClipSim	mean ClipSim	max ClipSim
Basic Label	0.14	0.48	**0.95**
Plain Triples	**0.18**	0.54	0.88
Verbalised Triples	0.17	0.55	0.92
DBpedia Abstract	0.16	**0.60**	0.92

Table 1: The CLIPScore is computed on the entire dataset consisting of the generated images of 1,500 fictional characters.

Findings about the CLIPscore

- No correlation with the # of relation and the # of distinct relation
- Not significantly impacted in the properties values
- Generally, low quality with characters with known visual representations; however, higher for fictional characters from novels.

Human evaluation

A human evaluation is conducted on generated images of randomly chosen 10 fictional characters presented to 101 people.

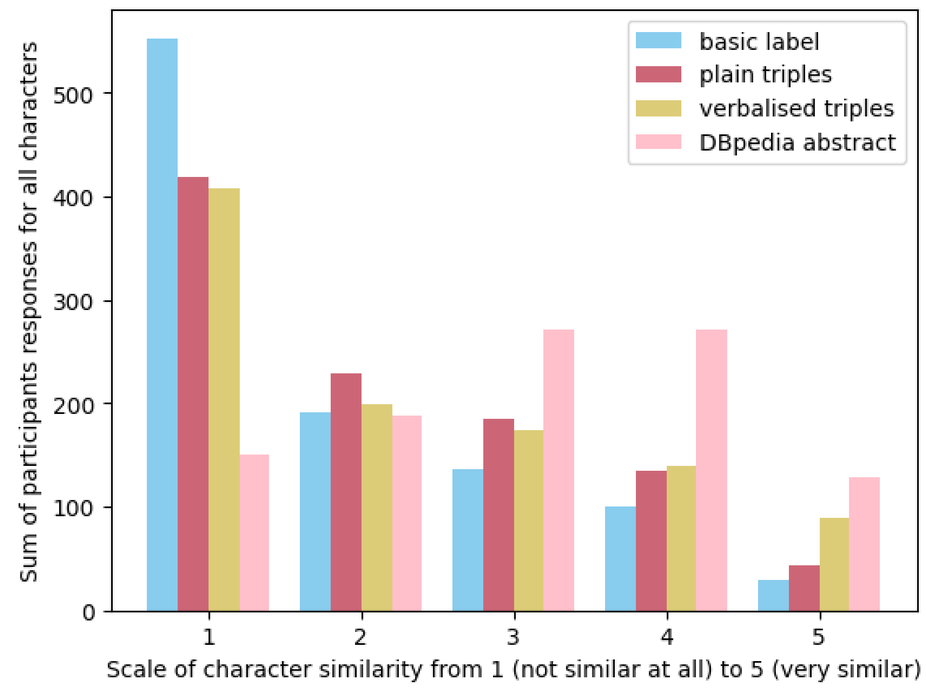

Figure 3: Distribution of the human evaluation survey results for all four prompt types.

- Human evaluation gives more similar results with CLIPScore rather than UQI and FID [5].
- Krippendorff's alpha equals 0.17.
- Key elements influencing the decision are mainly based on the physical characteristics of the character, and the coherence of it with its cultural context.

Limitations

- **Dataset:** contains photography, limited to English, prompt plain triples design choices
- **Generative model:** SD quality and bias issues; choice of negative prompts
- **Evaluation:** only based on entities small subset having an image; 1-5 score is not enough to evaluate the image quality
- **Metric:** the CLIPscore is not sensitive to the triples properties.

→To mitigate any copyright and/or privacy risks, we stress that our method is not suggested to be directly deployed into Wikidata, as we think that using AI-generated images might not be robust. Should this method be used for image completion, we encourage clearly watermarking images as AI-generated.

References

[1] Gabriel Amaral, Odinaldo Rodrigues, and Elena Simperl. WDV: A Broad Data Verbalisation Dataset Built from Wikidata. In Ulrike Sattler, Aidan Hogan, Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato, editors, *The Semantic Web – ISWC 2022*, pages 556–574, Cham, 2022. Springer International Publishing.

[2] Brümmer, Martin and Dojchinovski, Milan and Hellmann, Sebastian. DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3339–3343, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[3] Yixiong Chen. X-IQE: eXplainable Image Quality Evaluation for Text-to-Image Generation with Visual Large Language Models. *arXiv preprint arXiv:2305.10843*, 2023.

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2018.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[7] Domonkos Varga. Full-Reference Image Quality Assessment Based on an Optimal Linear Combination of Quality Measures Selected by Simulated Annealing. *Journal of Imaging*, 8(8), 2022.

[8] Denny Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledgebase . *Communications of the ACM*, 57(10):78–85, 2014.