

Experiment 1: Generating Harry Potter image sequence using Midjourney v6.1

Background: Since the 2024 AI Index Report, Midjourney v6.1 was released. In Section 2.4 (page 96) of the 2024 AI Index Report, a sequence of Harry Potter images are generated using each of the available Midjourney models under the prompt “a hyper-realistic image of Harry Potter.” In this section, I include the image generated by v6.1 using the same prompt.



Experiment 2: One-shot out-of-distribution generation

Background: The Midjourney models (v4 and later) were trained primarily on the LAION-5B dataset (<https://laion.ai/blog/laion-5b/>), which contains images up to around 2022. In this section, we prompt different Midjourney models to produce a copy of an out-of-distribution image that it has likely never seen in its training dataset, namely the Chill Guy meme (https://en.wikipedia.org/wiki/Chill_guy). This meme was popularized in October 2023 and was presumably never seen before then. Specifically, we use a one-shot prompt containing the Chill Guy meme image as a reference, and the prompt: “An exact replica of the Chill Guy meme (example attached).” We run this experiment across all available Midjourney models using default stylize, weirdness, and variety parameters.

Commentary: For v1, v2, and v3, the generated images are significantly lower-quality than the other images. The community agrees that v4 saw a significant improvement in image generation quality. The improvement starting at v4 may be attributed to the following: (1) The dataset used to train the model (LAION) was improved / updated to include more robust training data (2) The amount of compute available significantly increased, starting from v4, Midjourney was trained on Google TPUs (3) V4 was trained on a new architecture for over 9 months (4) more complex prompts in the training set allowed for better generalization and prompt alignment.

We also see that the image generated by v4 most closely aligns with the reference image, despite having non-coherent text towards the bottom of the image. Starting v5 and onwards, the community notes that Midjourney focused heavily on creating realistic human qualities in generated images. This roughly aligns with our empirical results as the images generated from v5, v5.1, v5.2, v6, and v6.1 all have stronger human-like qualities.

Midjourney v1



Midjourney v2



Midjourney v3



Midjourney v4



Midjourney v5



Midjourney v5.1



Midjourney v5.2



Midjourney v6



Midjourney v6.1



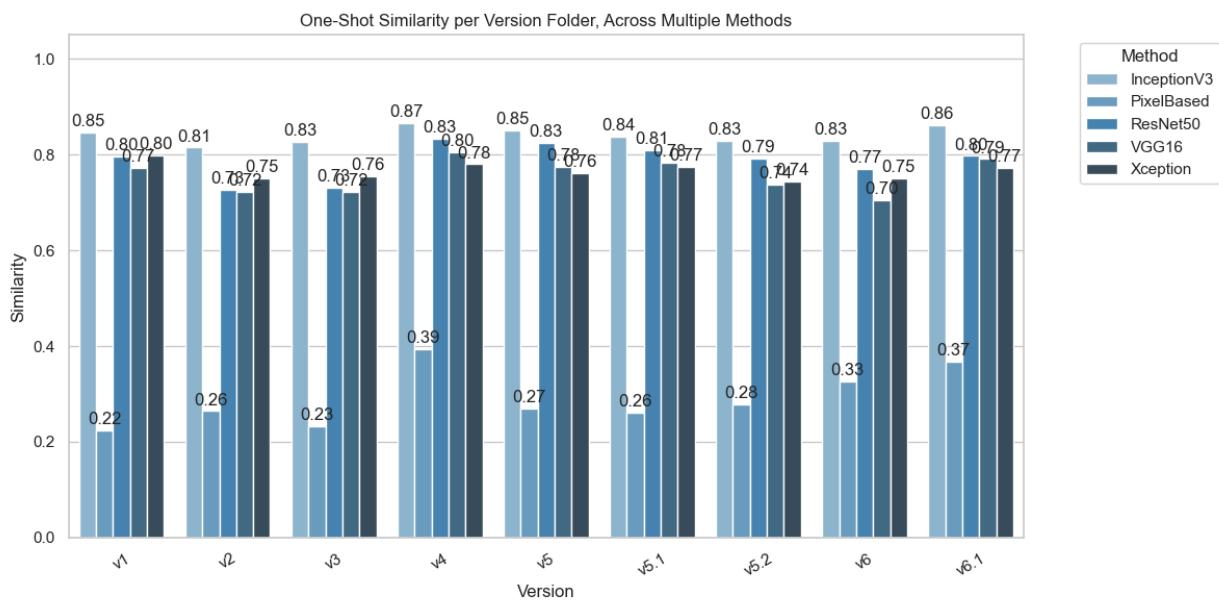
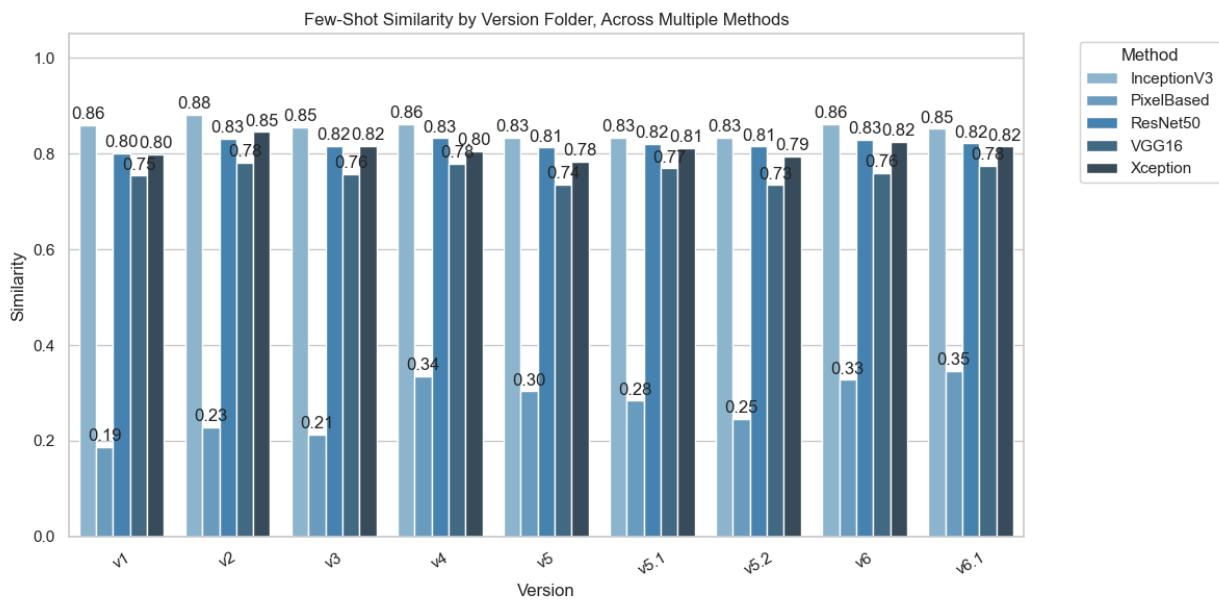
Experiment 3: One-shot and few-shot image similarity

Background: In this experiment, we provide the model with one reference image in the one-shot case, and five reference images in the few-shot case. We then use multiple methods to test for image similarity between the generated image and the reference image. We run this experiment across all Midjourney models from v1 to v6.1 and analyze the output quality. Specifically, the prompt used for the one-shot case is “A hyper-realistic image of Harry Potter similar to the image provided” and the prompt used for the few-shot case is “A hyper-realistic image of Harry Potter similar to the images provided.” We generate four images for each prompt to reduce noise.

Algorithm: To analyze image similarity, we use four methods: (1) pixel-based approach that uses structural similarity index measure (SSIM) to measure how visually similar the images are on a pixel-by-pixel basis after resizing to match image dimensions (2 - 4) CNN-based approach that loads a pretrained CNN, extracts a feature vector from a specific layer, and compares the feature vectors using cosine similarity. For methods (2

- 4), the similarity scores range from 0 (complete opposite) to 1 (identical) after mapping from the raw cosine similarity value of [-1 ... 1]. The similarity scores are then averaged across each of the four generated images.

The results of the one-shot and few-shot experiments are below. Note that v1, v2, and v3 noticeably underperform the other models. Interestingly, v4 has high performance across the board, with similar scores to Midjourney's state-of-the-art v6.1 model and in some cases surpassing it.



Experiment 4: Effect of parameter values for v6.1

Starting with v3, the stylization parameter was added, which Midjourney defines as: “Influences how strongly the Midjourney aesthetic is applied. Low stylization values produce images that closely match the prompt but are less artistic. High stylization values create images that are very artistic but less connected to the prompt.” The default value is 2500 for v3, and 100 for all subsequent models. Starting with v5, the weirdness parameter was added, which Midjourney defines as: “Introduces quirky qualities to your generated images, resulting in unique and unexpected outcomes.” The default value is 0 for all applicable models. All models have the variety parameter which Midjourney defines as: “Influences how varied the images are. High values will produce more unusual and unexpected results and compositions. Lower values have more reliable, repeatable results.” Additionally, there are three options—relax, fast, and turbo—for the generation speed for all models; the default is fast for all models.

I experimented with different values for these parameters, but did not notice much difference. For example, using the same prompt—“a hyper-realistic image of Harry Potter”—here is a comparison using v6.1 with different parameters. The difference is noticeable but not hugely significant. As we increase the stylization value, the facial details and background become more vivid and have greater variety. There is a more significant difference between the stylize 0 and stylize 100 images due to the facial details. However, the background remains relatively similar.

fast, stylize 0, weirdness 0, variety 0 (minimum stylize)



fast, stylize 100, weirdness 0, variety 0 (default)



fast, stylize 1000, weirdness 0, variety 0 (maximum stylize)



Now, fix the stylize parameter and vary the weirdness parameter. Again, we run three different generations. The maximum weirdness image has a distinct background and the average weirdness image shows Harry Potter as a younger person with different facial characteristics and features.

fast, stylize 100, weirdness 0, variety 0 (default, minimum weirdness)



fast, stylize 100, weirdness 1500, variety 0 (average weirdness)



fast, stylize 100, weirdness 1500, variety 0 (maximum weirdness)



Finally, fix stylize and weirdness and vary variety. To see the effect of the variety parameter, we generate 4 images per setting and compare the difference between the images.

fast, stylize 100, weirdness 0, variety 0 (default, minimum variety)



fast, stylize 100, weirdness 0, variety 50 (average variety)



fast, stylize 100, weirdness 0, variety 100 (maximum variety)

