Visualizing Semantic Walks

Shumeet Baluja
Google Research

David Marwood
Google Research

Abstract

An embedding space trained from both a large language model and vision model contains semantic aspects of both and provides connections between words, images, concepts, and styles. This paper visualizes characteristics and relationships in this semantic space. We traverse multi-step paths in a derived semantic graph to reveal hidden connections created from the immense amount of data used to create these models. We specifically examine these relationships in the domain of painters, their styles, and their subjects. Additionally, we present a novel, non-linear sampling technique to create informative visualization of semantic graph transitions.

1 Introduction & Semantic Similarity

The rapid advances in transformer-based [9] text-to-image systems [7, 8, 10, 5, 6] provides an opportunity to not only create new artwork, but also to examine the connection of artists and artistic styles in the data used to create the models. We present our first results towards examining these relationships by visualizing transitions in an implied semantic graph. To encourage wider adoption, we use only non-proprietary models from CLIP [4] to convert text into a semantic space embedding and Stable Diffusion (SD) [7] to convert the embeddings to images, and employ only a single GPU.

For this study, we examined text prompts of the form $prompt_{o,a}$ = "a painting of o by a", where o is a concrete noun (*e.g.* a bird, car, girl) chosen from a set of 500 [3], and a is a well known artist that is likely to be in the models. A fundamental assumption of this work is that similar artists are represented similarly in the semantic space. To verify this, we computed the average embedding for each artist, $C_A = \sum_{o \in Objects} CLIP(prompt_{o,a})/|o|$. We used a set of 100 famous painters [1] and manually inspected the similarity for several of the best known artists; three are shown in Table 1.

At least subjectively, the results support the assumption. Artists of a similar school or style are closer in the semantic space. Given this, we can synthesize a fully-connected "semantic graph," where the nodes ($prompt_{o,a}$) are connected by edges weighted by the distance between their endpoint embeddings. We created a graph using 500 objects and 16 artists to generate 8,000 prompt-nodes and 64×10^6 edges. Numerous embedding distances were explored; the simplest, L_2, was used.

Table 1: Artist and Three Most Similar Artists (from set of 100)

Claude Monet	$\Longrightarrow$	Pierre-Auguste Renoir, Edouard Manet, Vincent Van Gogh
Gustav Klimt	$\Longrightarrow$	Egon Schiele, Hilma af Klint, James Abbot McNeill Whistler
Michelangelo	$\Longrightarrow$	Leonardo da Vinci, Carvaggio, Raffaello Sanzio (Raphael)

2 The Semantic Graph

With the ultimate goal of visualizing paths in the graph, we first discuss visualizing individual edges. At each of the prompt nodes, we use CLIP to convert the prompt to an embedding and use Stable Diffusion to render the image for the prompt. In our first attempt to visualize the transition frames, we interpolated the text embeddings simply, using $Interp(blend, A, B) = (1-blend)*A+blend*B$ and sampled the intermediate blends. While it was not *a priori* known whether this would yield visually

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

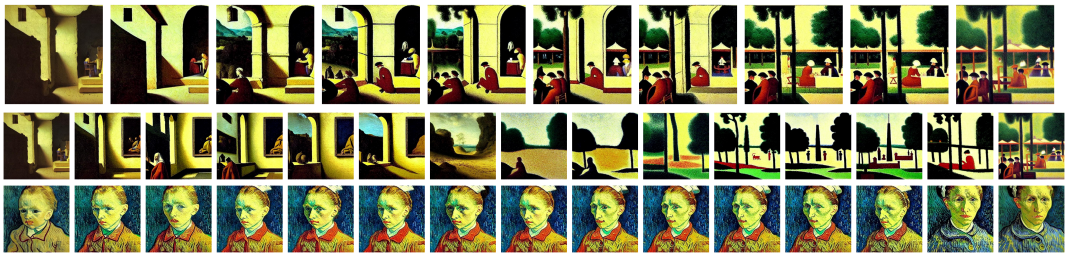

Figure 1: (top) A single edge: *a painting of a cave by Johannes Vermeer → a painting of a cafe by Georges Seurat*. (middle) Shortest path when the direct edge is removed visits 9 nodes, see text. (bottom) Baby → Grandmother by Van Gogh, the path traverses through 7 nodes: *a painting of {baby, kid, daughter, girl, woman, aunt, grandmother} by Van Gogh*.

interpretable results, we found consistent coherence across the vast majority of trials. However, there were three modifications required to make the visualization more compelling.

- Stable Diffusion has large basins of attraction when generating images. When linearly interpolating between two prompts, most frames showed very little visual change with (usually) a single large change — at the point where the basin of attraction 'flipped', typically near the middle. This led to uninteresting animations. Instead, we used a non-linear, adaptive, sampling procedure: initialized with frames rendered from the two endpoints, we computed the largest *visual* distance between consecutive frames (pixel difference) and rendered a new frame at the midpoint by mixing the *embedding* by the distance specified by the visual differences. The process was repeated until the desired number of frames were created. This over-sampled regions of largest visual change.
- There is no constraint that the two endpoints will have compatible images – *e.g.* similar regions of light and dark / colors. Starting with the same noise (by using the same random seed) provided sufficient bias to increase the likelihood of compatible images.
- For smoothness in intermediate images, we interpolated Stable Diffusion's image latents from the two endpoints using the same $Interp$ function. The interpolated latents (5%) were mixed with the diffusion initialization noise (95%); larger latent weight severely limited SD's progress.

Figure 1(top) shows interpolated frames from a single graph edge: *a painting of a cave by Johannes Vermeer → a painting of a cafe by Georges Seurat*. Let's now examine what happens in a non-degenerate case. In Figure 1(middle) the direct edge is removed and we find the shortest path [2]. To encourage paths with smaller hops, we can exponentiate the distances. The path visits 9 nodes:

painting of {cave,oil,paint,sand} by Vermeer → painting of {sand,earth,country,town,cafe} by Seurat

This path is a combination of both visual and text similarities. We can also shift the focus away from artists to visualize the relationships in subjects by holding the artist constant. For example, to move from *a painting of a baby by Van Gogh → a painting of a grandmother by Van Gogh*, the path travels through *{baby, kid,daughter,girl,woman,aunt,grandmother}* – a very intuitive progression; see Figure 1(bottom). Additionally, we can visualize the transitions between (dis)similar subjects, artists styles, and their combinations; many examples are given in the Supplementary Material.

3 Conclusions & Discussion

The semantic graph, as inferred by the semantic space embeddings, provides the basis for a new distance measurement of related concepts, artists, and artistic styles. CLIP, and similar embeddings, may be particularly well suited for this task as they are trained to combine enormous amounts of both visual and textual samples. By creating a graph to traverse, we can move through the semantic space easily. To facilitate the visualizations, we non-linearly sampled the walk along the edges; this is well suited to the discontinuities inherent in diffusion image generation. We also provided guidelines for enhancing visual smoothness. This work was done with publicly available models and modest computation. We hope that this work encourages continued combinations of graph analysis techniques with the new distance measurements in general, as well as specifically providing insights and relationships into the visual artistry of painters and into the creation of new artworks based on the combination of their skills.

References

[1] the artwolf. Most important western painters. https://theartwolf.com/most-important-painters/, 2022. accessed: 2022-9-15.

[2] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[3] english vocabs. 500 concrete nouns example list. https://englishvocabs.com/concrete-nouns/500-concrete-nouns-examples-list/, 2021. accessed: 2022-9-15.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[6] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv:2112.10752*, 2021.

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*, 2022.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

4 Supplementary Materials

Table 2: 16 Artists Examined

Claude Monet	Vincent Van Gogh	Pierre-Auguste Renoir	Albrecht Durer
Rembrandt Van Rijn	Leonardo Da Vinci	Hieronymus Bosch	Paul Gauguin
Gustav Klimt	Winslow Homer	Henri de Toulouse-Lautrec	Johannes Vermeer
Michelangelo Buonarroti	Gustave Courbet	Sandro Botticelli	Georges Seurat

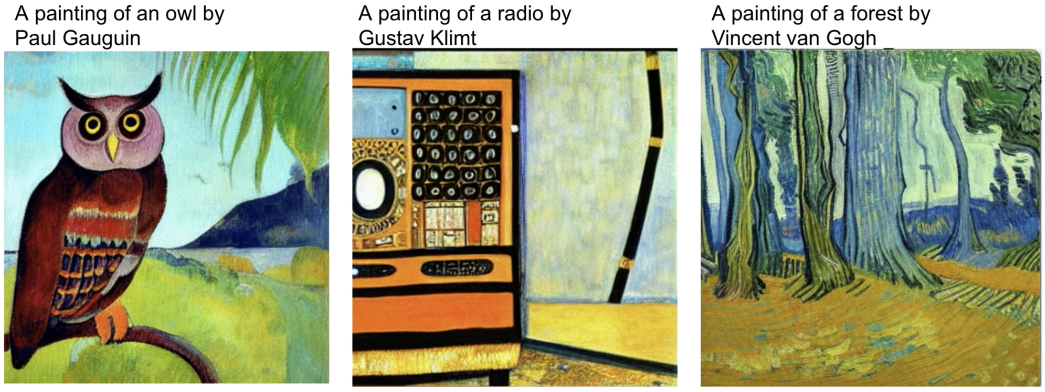

Figure 2: Examples of paintings created with Stable Diffusion. Rendering at 512x512 pixels. Each took approximately 20 seconds on a single P100.

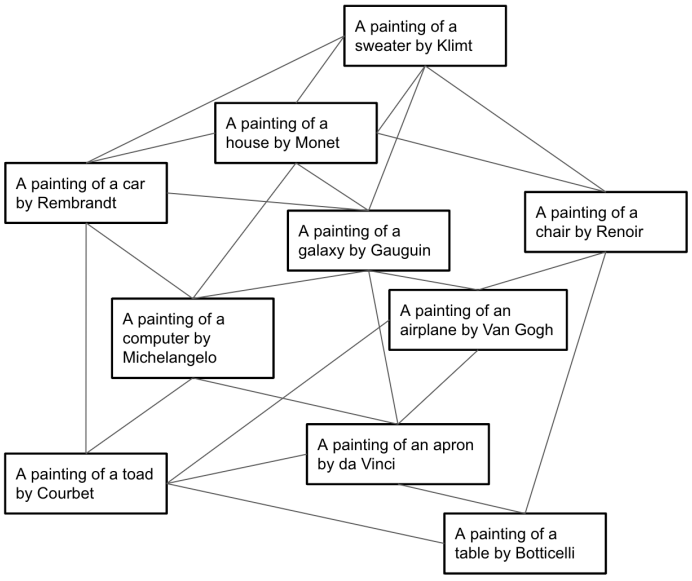

Figure 3: A small subset of the fully connected graph used. The full graph has 8,000 nodes with 64 million edges.

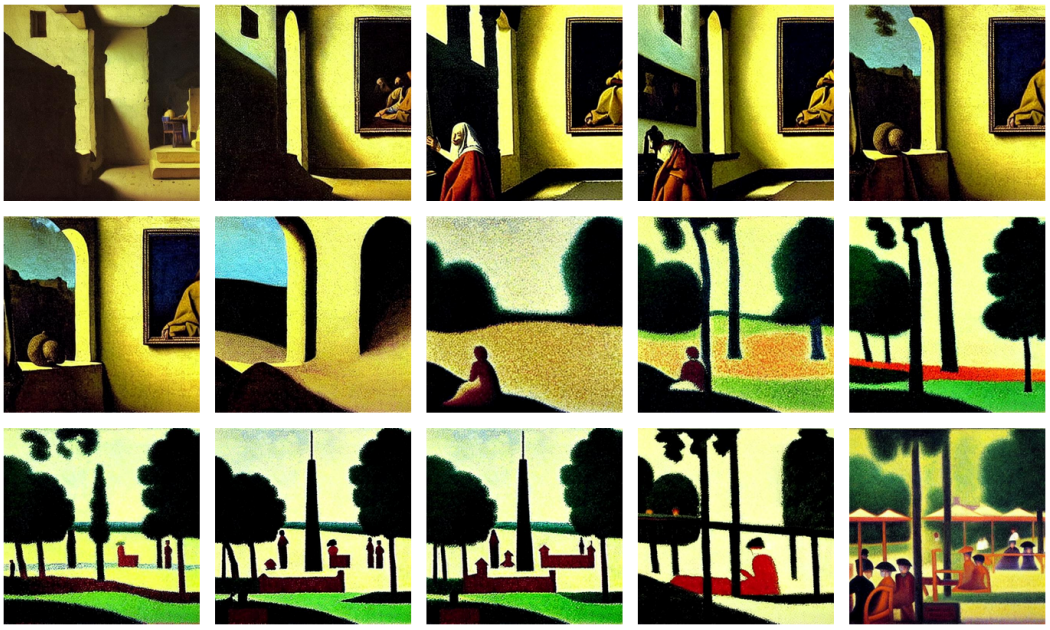

Figure 4: Enlargements from Figure 1. Path from *a painting of a cave by Johannes Vermeer* & *a painting of a cafe by Georges Seurat* when direct edge is removed goes through *a painting of {cave,oil,paint,sand} by Vermeer* to *a painting of {sand, earth,country,town,cafe} by Seurat*.

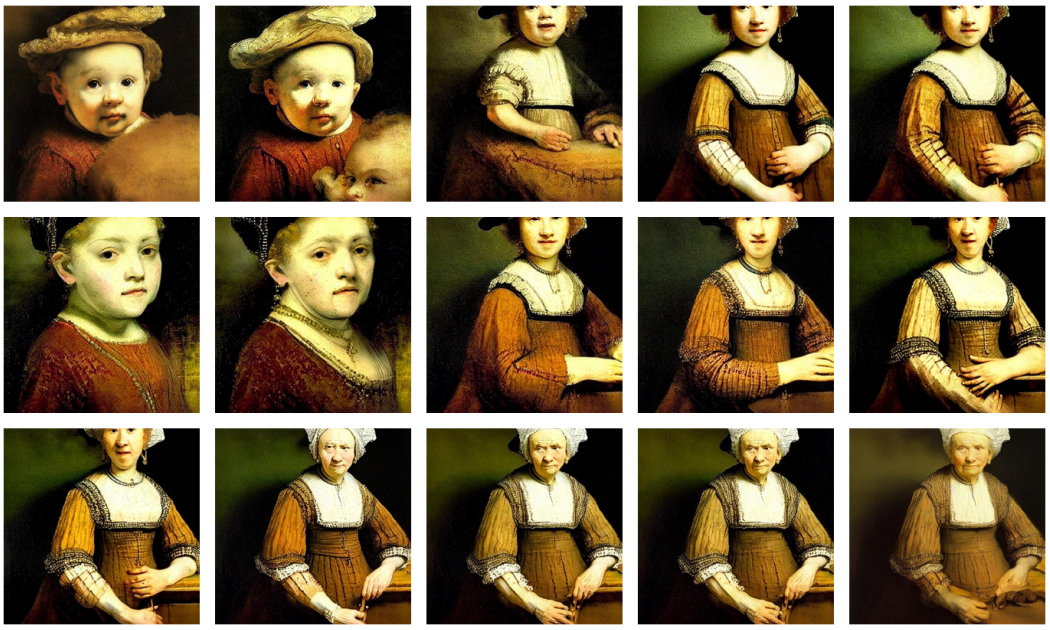

Figure 5: Reexamining transitions of Figure 1(bottom) with Rembrandt Van Rijn. The same path (respectively) is followed as with Van Gogh. *a painting of a Baby by Rembrandt Van Rijn* $\rightarrow$ *a painting of a Daughter by Rembrandt Van Rijn* $\rightarrow$ *a painting of a Girl by Rembrandt Van Rijn* $\rightarrow$ *a painting of a Woman by Rembrandt Van Rijn* $\rightarrow$ *a painting of an Aunt by Rembrandt Van Rijn* $\rightarrow$ *a painting of a Grandmother by Rembrandt Van Rijn*

Figure 6: Disparate Artists and disparate subjects. *a painting of Buttons by Paul Gauguin* to *a painting of a Casino by Henri de Toulouse-Lautrec*. Since the direct edge is removed, the path traverses the nodes: *a painting of Buttons by Paul Gauguin* $\rightarrow$ *a painting of a Coin by Paul Gauguin* $\rightarrow$ *a painting of an Orange by Paul Gauguin* $\rightarrow$ *a painting of a Fruit by Paul Gauguin* $\rightarrow$ *a painting of an Apple by Paul Gauguin* $\rightarrow$ *a painting of an Apple by Henri de Toulouse-Lautrec* $\rightarrow$ *a painting of a Fruit by Henri de Toulouse-Lautrec* $\rightarrow$ *a painting of an Orange by Henri de Toulouse-Lautrec* $\rightarrow$ *a painting of Oil by Henri de Toulouse-Lautrec* $\rightarrow$ *a painting of a Casino by Henri de Toulouse-Lautrec*.

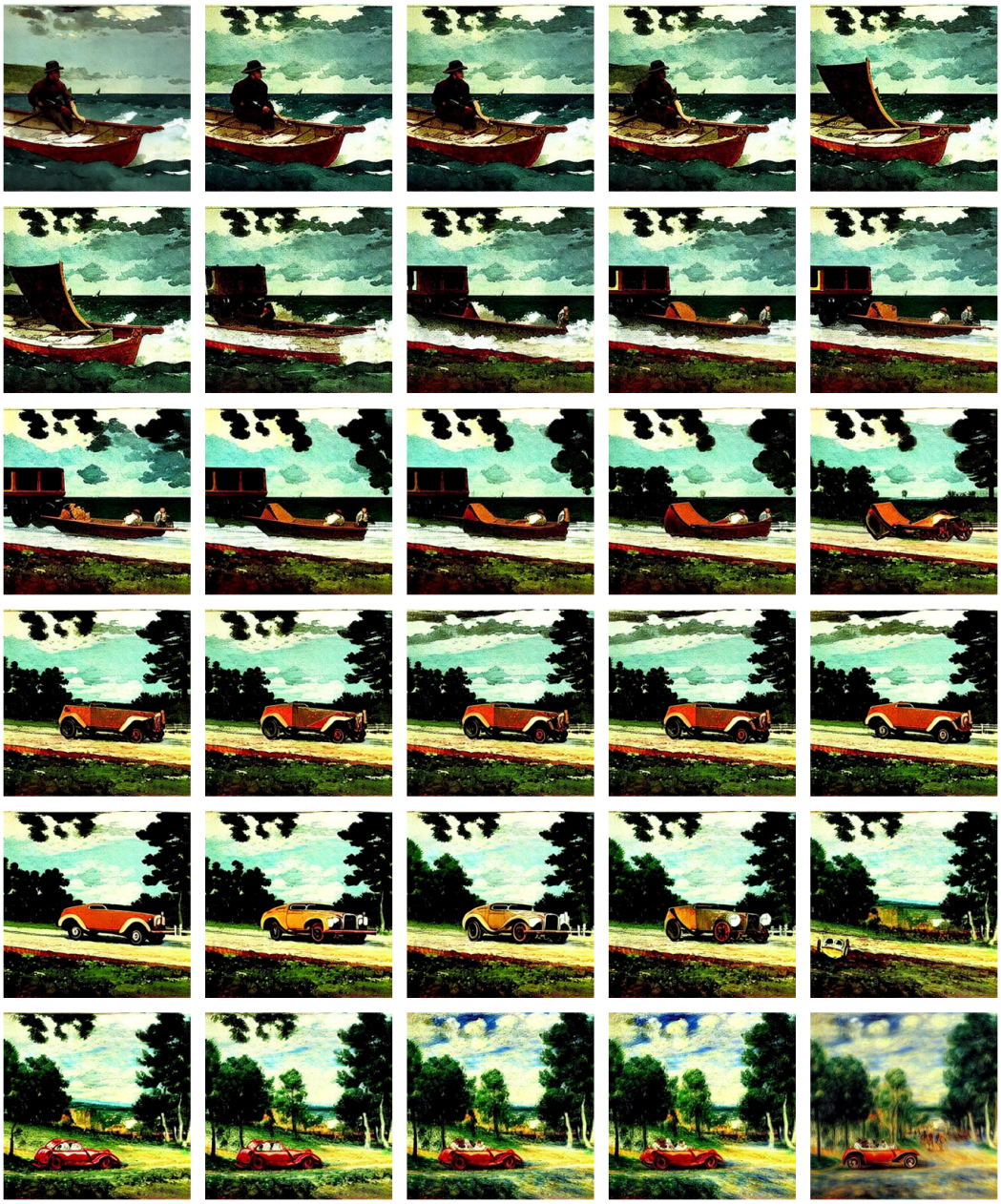

Figure 7: Disparate Artists and disparate subjects: Homer's Boat to Renoir's Car. a painting of a Boat by Winslow Homer $\rightarrow$ *a painting of a Vehicle by Winslow Homer* $\rightarrow$ *a painting of a Car by Winslow Homer* $\rightarrow$ *a painting of a Car by Pierre-Auguste Renoir*.

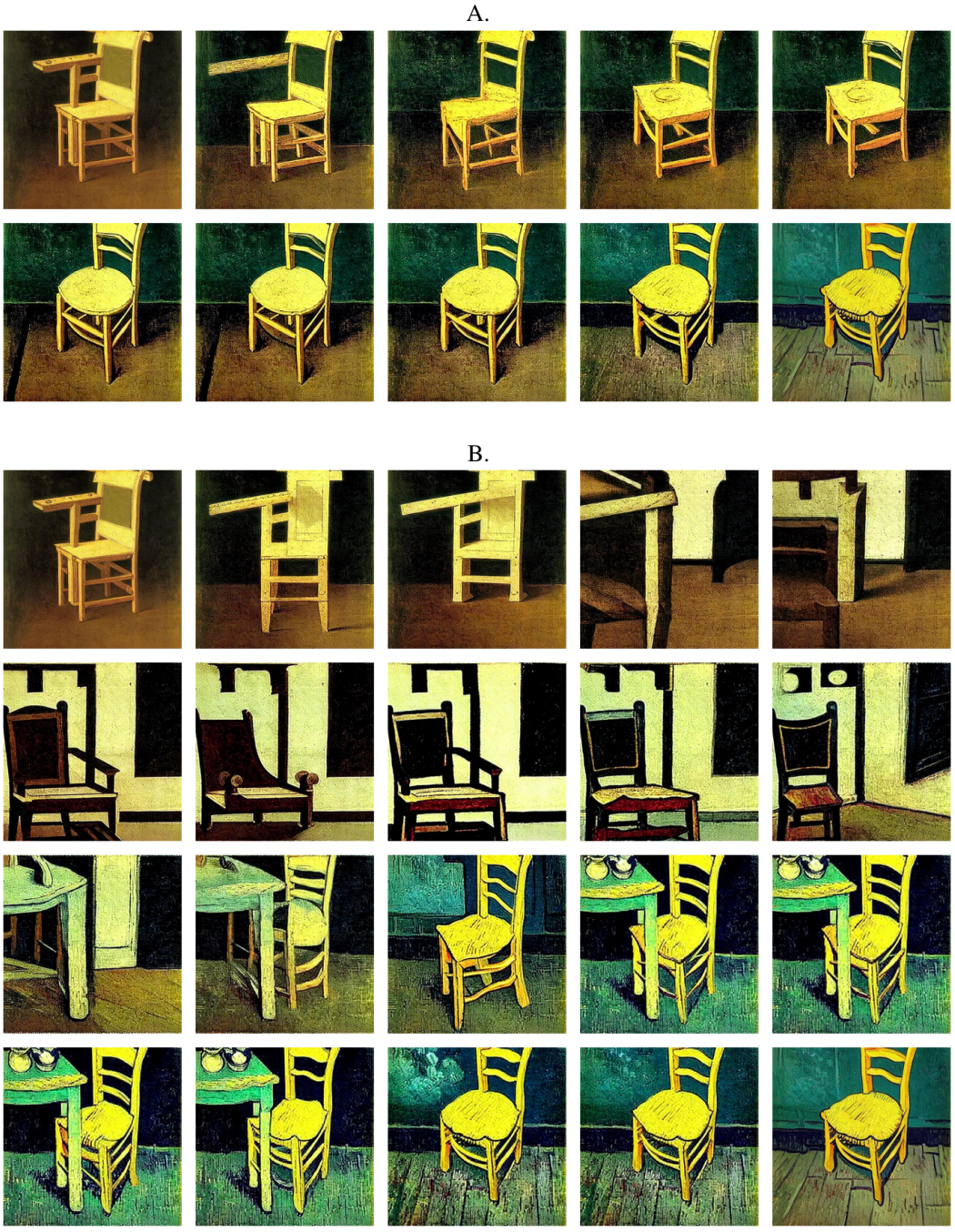

Figure 8: Examining how different artists paint the same object. A. Transition from *a painting of a chair by Leonardo Da Vinci* to *a painting of a chair by Vincent Van Gogh*. B. If the direct path is removed, the walk traverses the nodes: *a painting of an armchair by Leonardo Da Vinci*, *a painting of an armchair by Vincent Van Gogh*, *a painting of a chair by Vincent Van Gogh*.

Figure 9: A re-examination of *a painting of a cave by Johannes Vermeer* to *a painting of a cafe by Georges Seurat*. However, here we do not allow intermediate Vermeer or Seurat prompts; the path must go through another artist(s). The path traverses through the nodes: *a painting of a Cave by Johannes Vermeer* $\rightarrow$ *a painting of a Cave by Rembrandt Van Rijn* $\rightarrow$ *a painting of a Cave by Pierre-Auguste Renoir* $\rightarrow$ *a painting of a Gift by Pierre-Auguste Renoir* $\rightarrow$ *a painting of a Cafe by Pierre-Auguste Renoir* $\rightarrow$ *a painting of a Cafe by Georges Seurat*.

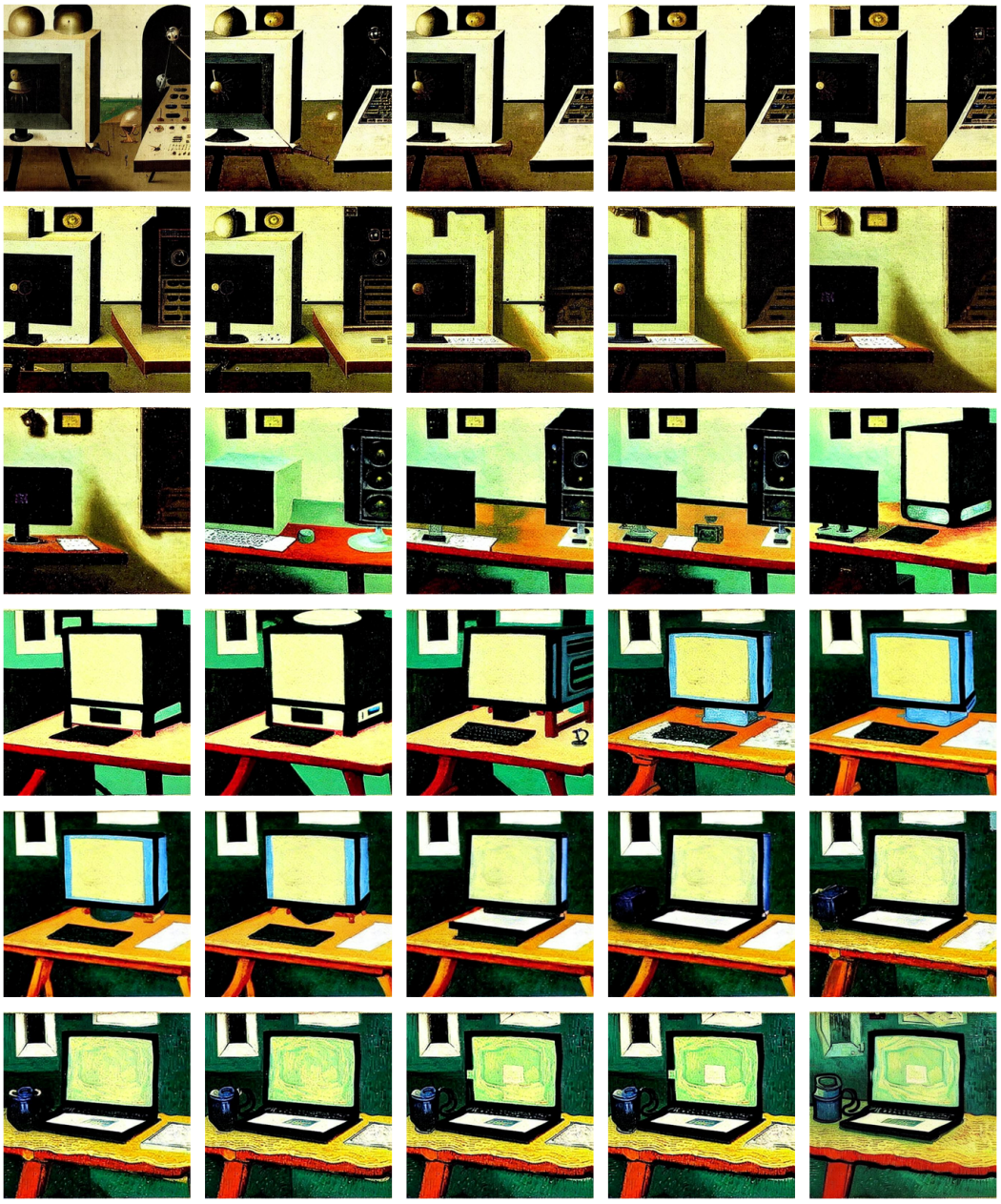

Figure 10: Unlikely Subjects and Impossible timelines, moving from *a painting of a Computer by Hieronymus Bosch* to *a painting of a Laptop by Vincent Van Gogh*. The path traverses: *a painting of a Computer by Hieronymus Bosch* $\rightarrow$ *a painting of a Computer by Rembrandt Van Rijn* $\rightarrow$ *a painting of a Computer by Vincent Van Gogh* $\rightarrow$ *a painting of a Laptop by Vincent Van Gogh*. Note: Hieronymous Bosch (1450-1516), Van Gogh (1853-1890).