
Gaudí: Conversational Interactions with Deep Representations to Generate Image Collections

Victor S. Bursztyn
Northwestern University
Evanston, IL 60201
v-bursztyn@u.northwestern.edu

Jennifer Healey
Adobe Research
San Jose, CA 95110
jehealey@adobe.com

Vishwa Vinay
Adobe Research
San Jose, CA 95110
vinay@adobe.com

1 Introduction

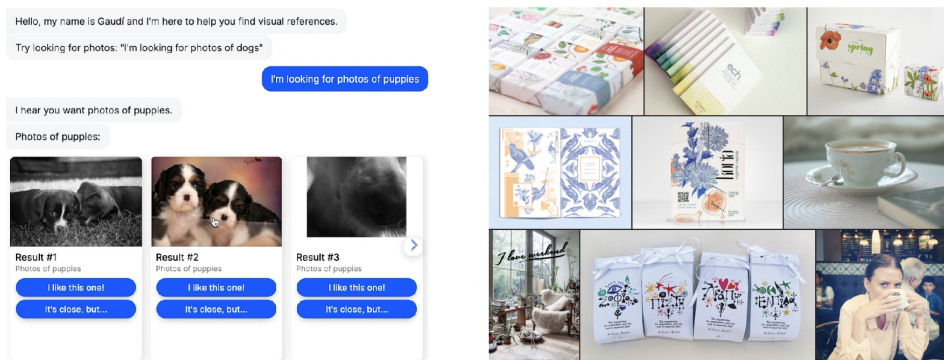


Figure 1: Left: Gaudí responding to a user query “I’m looking for photos of puppies.” Right: A mood-board created by a professional designer using Gaudí for the given project briefing: “You’re designing a new ecofriendly, highend coffee brand that is notorious for its floral flavors.” All images are from the BAM dataset [6].

Gaudí was developed to help designers search for inspirational images using natural language. In the early stages of the design process, designers will typically create thematic image collections called “mood-boards” (example shown in Fig. 1) in order to elicit and clarify a client’s preferred creative direction. Creating a mood-board involves sequential image searches which are currently performed using keywords or images. Gaudí transforms this process into a conversation where the user is gradually detailing the mood-board’s theme. This representation allows our AI to generate new search queries from scratch, straight from a project’s briefing, following a hypothesized mood.

Previous computational approaches to this process tend to oversimplify the decision space, seeking to define it by hard coded qualities like dominant color, saturation and brightness [3, 2]. Recent advances in realistic language modeling (e.g., with GPT-3 [1]) and cross-modal image retrieval (e.g., with CLIP [5]) now allow us to represent image collections in a much richer semantic space, acknowledging richer variation in the *stories designers tell* when presenting a creative direction to a client.

2 Methods

In this section, we present the following methods: image retrieval based on text only; image retrieval based on both a reference image and a text query (or simply “composed image retrieval” [4]); and mood-board generation by (i) using a project briefing to generate a natural language story, and (ii) using each step of this story as a text query for image retrieval.

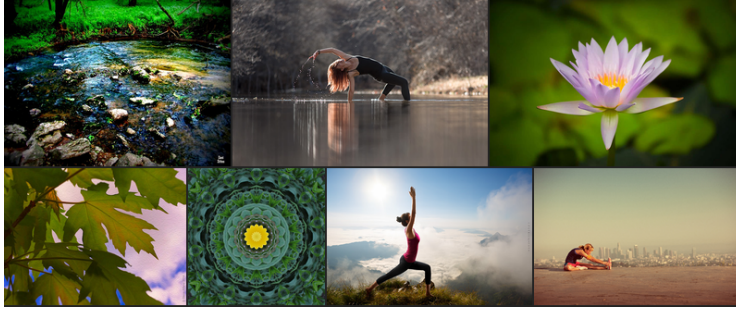


Figure 2: An automatically generated mood-board for the new project briefing: “You’re designing a new yoga kit for a highend company that is famous for its athletic clothes.” Images from BAM [6].

Method #1: Let q be a text query (e.g., “I’m looking for photos of puppies”) and Φ_q its cross-modal CLIP embedding. Let D be our image dataset and i an image $i \in D$, then Φ_i denotes the cross-modal CLIP embedding of i . The pairwise similarity between q and i can be denoted by $Sim(q, i) = \cos(\Phi_q, \Phi_i)$ such that text-based image retrieval can be defined as $Retr(q, D) = \operatorname{argmax}_{i \in D} Sim(q, i)$.

Method #2: Let q_m be a multi-modal query combining a reference image r (e.g., a previously selected puppy photo) with a text query q (e.g., “I want it more cheerful”) such that $q_m = \Phi_r \oplus \Phi_q$. To account for the extended shape of q_m and due to the cross-modal nature of CLIP, we generate an extended embedding Φ_i^{ext} for each image $i \in D$ defined as $\Phi_i^{ext} = \Phi_i \oplus \Phi_i$. Again, the pairwise similarity between q_m and i can be denoted by $Sim(q_m, i) = \cos(q_m, \Phi_i^{ext})$ such that composed image retrieval can be defined as $Retr(q_m, D) = \operatorname{argmax}_{i \in D} Sim(q_m, i)$.

Method #3: Let b be a project briefing expressed in natural language and Q a set of queries made by a professional designer when creating a mood-board. We condition GPT-3 with one fully detailed example (single-shot) of $b \rightarrow Q$, as described below, plus a new briefing for which we want to generate search queries. We use the DaVinci model with temperature = 0.7, top p = 1.0, response length = 80, and no penalties when sampling the output. Each generated query serves as input to method #1 or #2 in order to generate a mood-board as seen in Fig. 2.

GPT3 Input :

You're designing a new ecofriendly, highend coffee brand that is notorious for its floral flavors. =>
I'm looking for photos of women sipping coffee. I'm looking for photos of joyful coffee packages.
I'm looking for photos of coffee cups and books. I'm looking for photos of luxury coffee shops with plants.
I'm looking for images of floral packaging. I'm looking for images of floral packaging that seems a bit more craft.
I'm looking for images of blue, floral packaging that seems a bit more craft.
I'm looking for images of classy, colored, craft packaging. I'm looking for images of posters with blue birds and flowers.
I'm looking for images of posters with blue birds and flowers.
You're designing a new yoga kit for a highend company that is famous for its athletic clothes. =>

GPT3 Output :

I'm looking for photos of trees and grass. I'm looking for photos of water.
I'm looking for photos of leaves. I'm looking for images of women practicing yoga.
I'm looking for images of women practicing yoga and wearing athletic clothes.
I'm looking for images of women practicing yoga and wearing athletic clothes in nature.
I'm looking for images of women practicing yoga in nature.

3 Results

We recruited a professional designer for an in-depth exploration of Gaudí, leading to the ground-truth in method #3 and expert assessments of the automatically generated mood-boards (e.g., Fig. 2). The subject was very satisfied (5 in a 5-point scale) with the responsiveness of methods #1 and #2 afforded by CLIP. When asked about the quality of the generated mood-boards, the subject rated 4.5 (of 5) for the queries generated by GPT-3 and 3 (of 5) for the mood-boards. Besides a positive surprise with the queries, the subject suggested: “Although I would have picked different images, I see a story. This may be useful to marketers that use mood-boards in their work but are not used to crafting them.”

References

- [1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [2] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. May ai? design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.
- [3] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E MacKay. Imagesense: An intelligent collaborative ideation tool to support diverse human-computer partnerships. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, 2020.
- [4] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. *arXiv preprint arXiv:2108.04024*, 2021.
- [5] 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. *arXiv preprint arXiv:2103.00020*, 2021.
- [6] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE international conference on computer vision*, pages 1202–1211, 2017.