BehaviorGPT for Visual Art: A Foundation Model for Aesthetics

Rickard Brüel Gabrielsson,* **Vasudev Gupta, et al.**
Unbox AI

Abstract

We introduce BehaviorGPT-v3, the first behavioral foundation model for visual art and aesthetics. Trained on 215 billion human interactions (4.7 trillion tokens) across major art and design platforms, BehaviorGPT is a 0.5B-parameter Transformer model that learns semantic meaning not from pixels or text-image pairs but from sequences of user behavior. By treating aesthetic exploration as a language of actions, the model captures intent, contextual preference, and non-verbal nuance. In large-scale deployments, BehaviorGPT significantly outperforms leading search and recommendation systems: +16% in search conversion, +24% in recommendation performance, +11% in dynamic categorization, and +14% in SEO and assortment optimization, while requiring 12x fewer human resources. We also demonstrate qualitative gains in personalization, natural language navigation, and motif generation. Retraining yields 2.5% annual compounding gains, underscoring the long-term value of behavioral data loops. We position BehaviorGPT in contrast to vision foundation models (VFMs), emphasizing its ability to model behavior-first semantics, and conclude with ethical considerations around attribution, transparency, and augmentation of human creativity.

1 Introduction

Foundation models are large neural networks trained on massive datasets to learn general-purpose representations. They have shown remarkable success in domains such as language, vision, genomics, and behavior [1], [2]. Their effectiveness lies in capturing complex patterns from diverse data before fine-tuning or prompting for specific tasks.

Figure 1: 3D projection of art motifs using BehaviorGPT

Most foundation models revolve around languages of tokens: human language, genomics/DNA, health records, and so on. In this paper, we focus on the non-verbal language of visual art and aesthetics. We are not discussing image generation (e.g., diffusion models) or CLIP-like models, but the language spoken by human actions upon art—sequences of behaviors around art objects and aesthetic preferences. We treat these as a language to uncover what it reveals.

This domain offers unique opportunities to advance foundation models. Here are the key angles:

- *Behavioral semantics over pixel co-occurrences & object recognition:* Meaning is defined by statistical co-occurrences of motifs in user sequences, rather than pixels within an image

*Corresponding author: rickard@unboxai.com

Original post (published July 15, 2025): `https://research.unboxai.com/behaviorgpt-visual-art-and-aesthetics.html`

or across a static sample set. This grounds aesthetics in human interactions, revealing preferences that pixel-based models, sampling models, or scene understanding models (e.g., VFMs) miss. Indeed, current VFMs are all about recognizing object categories in the wild, while ours focuses on behavioral nuances. Creating manual datasets by asking users to describe visual art [3], [4] does not capture behavior; it captures small contexts of rationalizing, while actions speak louder than words.

- *Deep contextual and non-verbal nuances:* Art and aesthetics are inherently more contextual and non-verbal than textual data. Our extreme contextual approach uncovers deeper semantics that differ from merely extracting abstract concepts and information from an image for subsequent LLM processing, which is the focus of VFMs.

- *Bridging natural language and non-verbal navigation:* As a fundamentally non-verbal medium, visual art poses challenges in using human language for exploration and optimization (i.e., contextualizing language based on visual art). Mastering this enables personalized interfaces, echoing recent ethical discussions on AI augmenting creative discovery.

Indeed, visual art and aesthetics provide a great setting to elucidate the difference between abstraction vs. intent, global vs. local contexts, and the surprising depth of behavioral intelligence.

2 Global vs. contextual, abstraction vs. intent

"Mouse." Immediately, you have a picture in your mind—an example of what a mouse is. It happens instantaneously and unconsciously. However, if I say "computer mouse," you might get a different picture. If I had been discussing computers and then mentioned "mouse," the same shift occurs. This contextual interpretation of concepts is fundamental to our navigation and understanding of the world. Of course, it is difficult to know exactly what is meant by "mouse" without context, but as a user, you expect the system or AI to always understand your intent, which might differ from others.

Imagine entering "mouse" into a visual art platform: Do you mean the animal, a cartoon, a computer mouse, or the artist Stanley Mouse? Perhaps you imagine "cat and mouse" motifs, so we should show cat-heavy motifs by association?

These simplified examples highlight the one-size-fits-all challenge of intelligence and the difficulty in understanding users' intents. We need contextual cues about the user and their behavioral intent to deliver a tailored experience. When it works flawlessly, users are not grateful—they get what they expect. When it fails, frustration ensues, though the same response might be spot-on for someone else.

Figure 2: "Cat playing with a mouse" (left) is contextually clear and we all know how to interpret it because cat and mouse appear together a lot. "Dog playing with a mouse" is not as clear and well-defined, thus even the generative AI model generating image (right) gets confused.

Context is hierarchical and multi-faceted. Context can be that of a certain art-site, where users show a certain average behavior that is different from other sites e.g. at this site color might be more important than the object in the motifs, but that might be reversed at other sites or markets.

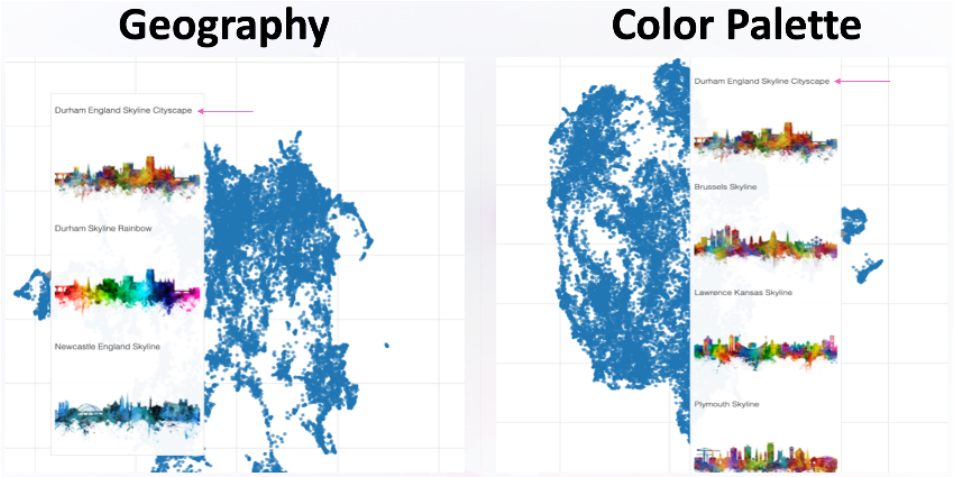

Figure 3: What is more important depends on context. What should define similarity: Geography or Color Palette? It depends on the context!

Ultimately, context is personal and you want to have a deep tailored experience so every user gets what they are looking for, and equally important, so the AI understands all the nuances of the content, i.e. the deepest possible intelligence.

It is interesting that VFMs around robotics are starting to think about context as a sequence of images (video) to help contextualize the current scene. This is what BehaviorGPT has been doing all along for human behavior.

3 The future of AI, intelligence, and diminishing returns

If it is not already clear, understanding the motifs and products that are interesting to a user is very contextual, and any such intelligence must be able to handle an infinite space of possible users and behaviors. Just something as simple as showing results for a query "stars" is extremely difficult, as the optimal results are unique for every single user, especially if the space of possible results is large. Indeed, even if one is expecting diminishing returns on recommendations and searching for a fixed assortment, the insights can get much deeper. For example, what products should be introduced? What should they be called and described? And can we create visual art from scratch that is tailored and personalized to a user on a pixel level? Indeed, generating an image just based on a prompt is not intelligent enough; we need to incorporate behavior to give the necessary contextual clues to optimally serve the user.

Recent VFMs emphasize multi-modal unification and information extracting, but behavioral grounding remains key for aesthetics.

4 CLIP is not enough

One of the most popular image and text models is CLIP (Contrastive Language-Image Pre-training). It uses self-supervised learning and contrastive learning to learn representations of images and text that can be used to compare within images and text separately and across. I.e., given a text, what images are most similar to that text, or vice versa, as well as given an image, what image is most semantically similar to that image. The goal was to enable zero-shot classification: Take a classification problem, describe each class with natural language, embed both the classification description and the image, and pick the classification that is closest in the embedding space to the image. The model might not

have been trained on this classification specifically, but if it has been trained on a lot of data and learned to generalize, it might still perform well on this task.

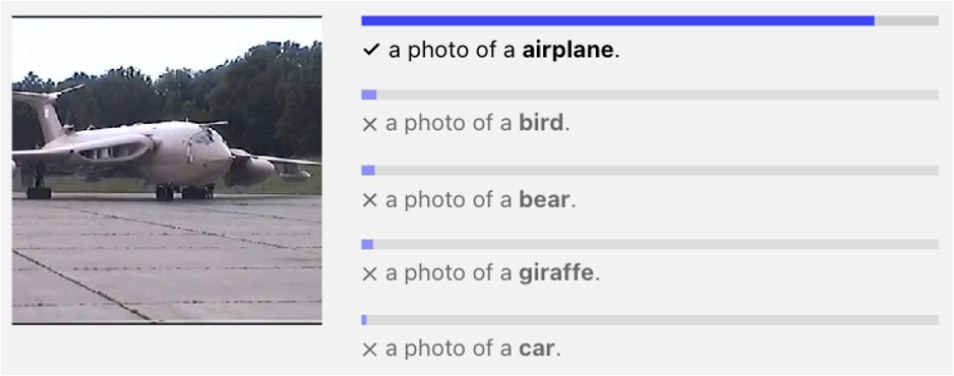

Figure 4: CLIP zero-shot classification.

As described, since CLIP embeds text and images in the same space, it can be used for both search (text-to-image retrieval) and semantically similar recommendations (image-to-image retrieval). However, there are a few fundamental limitations:

CLIP is not able to handle the context of a sequence or collection of images or texts. We want to be able to provide a sequence of user actions that are in the format of images (and text) and use that as context to the model to better serve the user.

The semantic similarity is not based on any strong behavioral signal. It's a set of datasets of text and image pairs from publicly available images. At most, it can aim at entailing some average text-image correspondences. However, it does not capture behavioral nuances. It will tell you that location is more important than color palette, based on text descriptions, rather than by how people act on the motifs (e.g., buy or view them).

CLIP is amazing, but it is not an autoregressive causal model, meaning we cannot ask it to generate text and images for us, or to incorporate or predict other actions such as visual art purchases or prices.

Post-CLIP advancements improve multi-modal aesthetics perception. Yet, they still undervalue and almost entirely ignore behavioral sequences for intent modeling.

Model	Flexible Language Prompts	Handles Sequences	Behavioral Grounding	Generate	Personalize
CLIP	✓	✗	✗	✗	✗
BehaviorGPT	✓	✓	✓	✓	✓

Table 1: Comparison of CLIP and BehaviorGPT across core capabilities.

5 How to learn the behavior of visual arts

We solve this and build a much deeper intelligence around visual arts and aesthetics by focusing on behavior and modeling a much more nuanced and interesting distribution—not text-image pairs, but text-image sequences that reflect user behaviors. So, instead of learning to map between image and text pairs, we have pairs of histories, and we learn to map which future corresponds to what history. Specifically, we have sequence-rich behavior, as shown in Figure 5. As in Figure 6, we remove the last action and train the model to predict it based on the previous actions in that sequence. This is a flexible way of modeling behavior.

This now allows the model to learn contexts and how to contextualize users and behavior conditioned on previous behaviors and contexts—for example, what are the optimal results if a user is in Boston

or if the first interaction has some particular motifs? It allows learning to semantically embed and understand images based on actual user behavior, and it can capture long dependencies—how different genres of visual motifs correlate with other distant genres. We can now also incorporate other types of data that might appear in these behavior sequences, predict them, and even generate as well.

Figure 5: Sequences of behaviors around visual art and aesthetics. Example of training data.

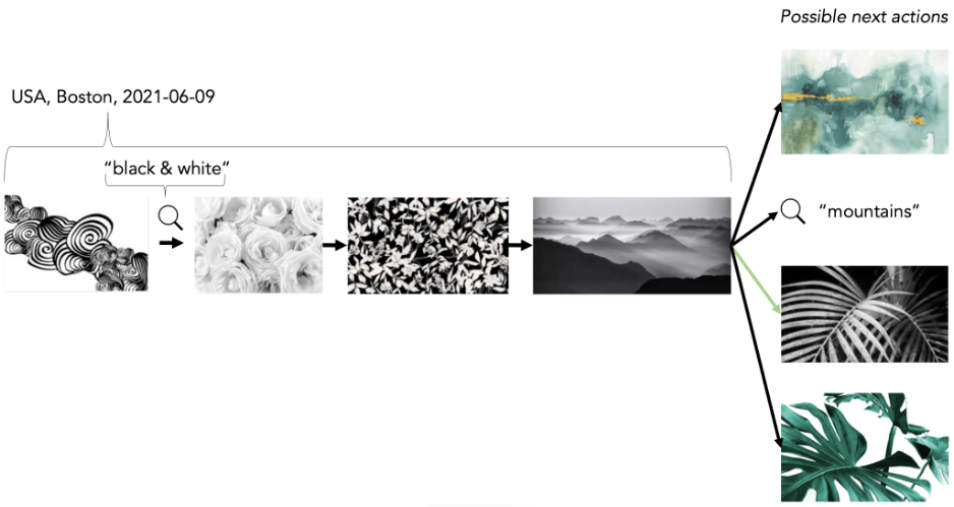

Figure 6: Learning to map from a behavioral sequence to the next action, picking the right one among a set of possible next actions—instead of the text-image pairs of CLIP. A powerful and flexible way of modeling behavior.

To accomplish this, we leverage CLIP as well as train a large Transformer to learn to contextualize and build on top of the CLIP embeddings (see Model section).

5.1 Data

We have rich behavioral data from a leading seller of art, posters, wallpapers, and canvases. This includes what users have viewed, purchased, and searched for, as well as demographics such as location and date/time. We create sequences as can be seen in Figure 5. We also have information about each visual artwork, such as the designer.

5.2 Compute

We trained on 96 H100s over two months.

5.3 Model

We use CLIP embeddings for the image and text inputs as the embedder, which are then fed into a contextualizer Transformer (see previous post) for details and Figure 7 for a simplified diagram). The CLIP provides non-contextual embeddings that serve as a good starting point for the model

to understand basic semantics, and the Transformer can then learn to contextualize a sequence of interactions by transforming the space, such that each user gets a unique, bespoke interpretation and metric space over visual art.

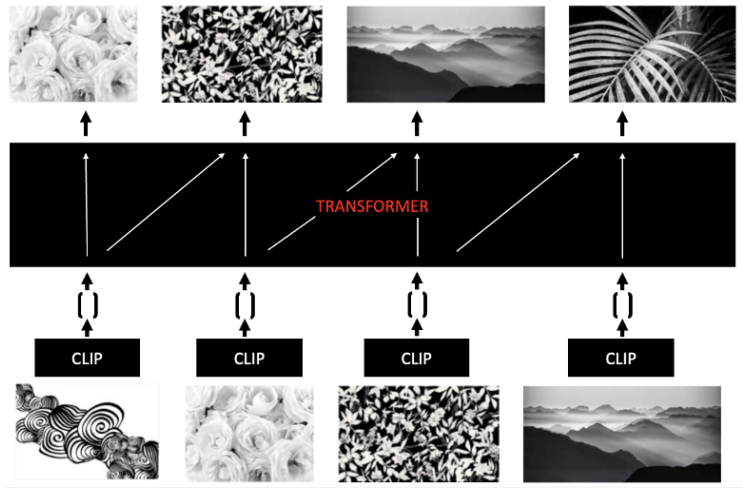

Figure 7: Simplified diagram of pushing CLIP embeddings through a Transformer that provides context and flexibility

5.4 Data sparsity & regularization (not forgetting)

The space of images is huge, if not infinite. Thus, learning a powerful and generalizing model on images is challenging. At the time, we did not have sufficient data to learn to understand all images, so we used CLIP to get embeddings for images that we trained a Transformer to contextualize. Now, since we train only on a small sample of all possible images, we don't want to overfit too much to those, as we want to be able to handle new motifs and art that is introduced and does not exist in the training data. One way of phrasing this is that we don't want to forget the abilities that CLIP has in being able to embed a broad set of images, though not contextualize them. To accomplish this, we introduced a special regularization. Roughly, it consists simply of the L2 distance between the input CLIP embedding and the output hidden embedding. Since the target matrix was produced by the CLIP embeddings, this allows us to default toward basic CLIP model behavior when the model is unsure of how to proceed.

6 Results

We now describe both the quantitative and qualitative results that BehaviorGPT for art was able to produce. See Table 2 for a summary.

Capability	Performance Uplift	Benefit
Search Conversion	+16%	More relevant and contextual search results
Recommendation Performance	+24%	Higher conversion through personalized suggestions
Dynamic Categorization	+11%	Smarter motif grouping and ranking
SEO & Assortment Optimization	+14%	Data-driven inventory expansion and content alignment
Personnel Efficiency	12x fewer required	Scalable performance with reduced human overhead
Annual Retraining Gains	+2.5% per year	Compounding intelligence from behavioral data loops

Table 2: Summary of quantitative and qualitative improvements of BehaviorGPT for art.

6.1 Quantitative

We evaluate the model on multiple different downstream tasks. Note that we have one single model trained flexibly that can be used for a wide range of downstream tasks and even business intelligence.

6.1.1 Search

Search is simply defined as the prediction of the next motif given a behavioral sequence that ends with a search query. We A/B tested this against a leading solution from Voyado, and we increased conversion on search by +16%. We also showed other desirable results:

- People spend longer time on the site when using our search (3:17 vs. 4:03, or +23%).
- Longer average search depth in exploring search results (2:19 vs. 2:64, or +21%).
- Fewer customers uploaded their own custom art and motifs (4.41% vs. 1.29%, or 3.4 times fewer custom uploads), suggesting our search helped them find what they were looking for more often.
- 20% fewer search refinements when using our search.
- +5% in average order value.

6.1.2 Recommendations

The same model was used for suggesting other motifs below an existing motif, which for the model is just recommending the next motif given a user sequence of actions that ends with a view-item of a motif. Our model increased conversion in an A/B test by +24% compared to a leading recommendation engine specializing in images and visual art.

6.1.3 Dynamic categorization

Since BehaviorGPT has learned a wide collection of actions, we can also ask the model what category of motifs to show next, what should be included in that category, and how they should be ranked. A/B tested against a market leader, our model increased conversion by +11%. Given that this was a site for visual art, exploration was a big part of the user experience, and users relied heavily on the categorization to click around and explore products—meaning it was an important uplift for the site.

6.1.4 SEO and assortment optimization

Since the model has learned to understand visual art and aesthetics generally and not merely the motifs and products that were available on the site, we could ask questions about what is not in the assortment but perhaps should be. This was work done by a big group of assortment specialists, but our model was able to fill gaps in the assortment to an estimated +14% increase in sales. This was achieved by asking both what motifs and what search queries users were looking for on the site but that were not part of the existing assortment.

6.1.5 Effects of retraining

Now we often talk about how better solutions (like a search, recommender, or assortment) lead to better data. Indeed, a big challenge is to help the model learn to understand optimal behavior from suboptimal behavior. Users' behavioral sequences are restricted and conditioned on the existing solutions that they are using to navigate. A clicked or even purchased product might not be the optimal product that they would consider if they had been shown more appropriate or personalized content. We work hard to enable us to learn this directly. Still, better solutions give better data, that in turn gives better solutions, and so on—i.e., intelligence compounds through data. We did a big retraining of the model every 12 months and saw that, on average, the model performance, A/B tested against the previous model, increased conversion by an average of 2.5% across tasks. This is like the interest rate, and an indication of the time-value of AI if you're not building foundation models on it.

6.2 Qualitative

First, we want to see how the model has learned to contextualize behavior around visual arts and aesthetics. See Figure 8 for a simple but still impressive example. If a user without any history

searches for “stars,” it gets photos of starry skies (more generic), but if a user has in their history searched for “oil painting” and interacted with a painting, and then searches for “stars,” the user gets oil paintings of starry skies. Of course, the model contextualizes the experience at every level; these are just the most obvious ones. However, manually mapping up these results would be impossible, as only with a history of three events and a 100K action space (searches and motifs), there are $100{,}000^3 = 1$ quadrillion such mappings to go through.

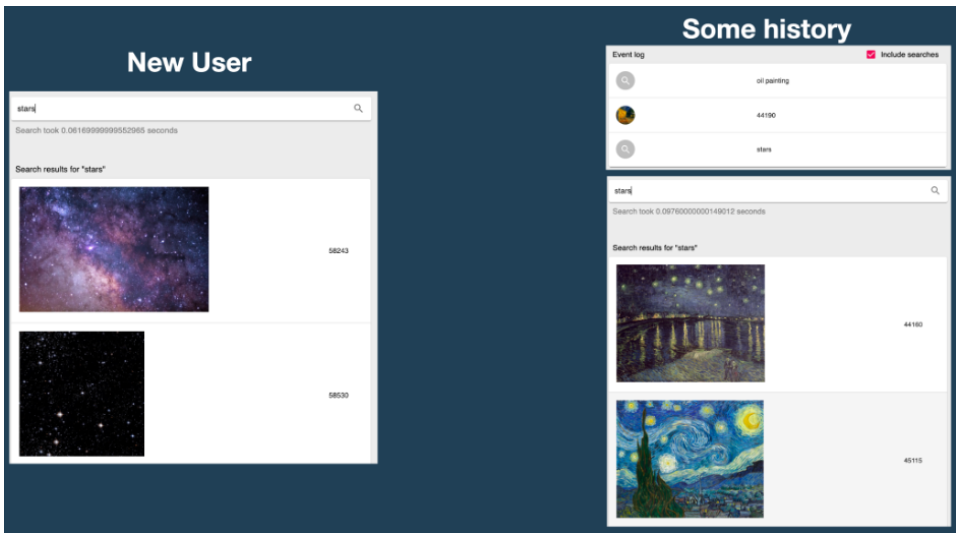

Figure 8: Behavioral and contextual understanding. With no history, a user (left) gets generic photos of starry skies. If the user (right) has searched for “oil painting” and interacted with one, the user gets oil paintings of starry skies.

It is also easy to see how the model learns to contextualize to user intention rather than “reasonable” abstraction. On this site, when people search for Elvis, they are not on average looking for only Elvis motifs, but they are looking for the aesthetics of Elvis. In Figure 9, you can see that without history, users also get Beatles motifs when searching for Elvis, because this is what the user data shows they want, while a non-contextual CLIP model only gives Elvis motifs. The latter might seem more “reasonable,” but it has much lower conversion. Again, it’s about modeling user intent, not our own expectations or idealized abstractions.

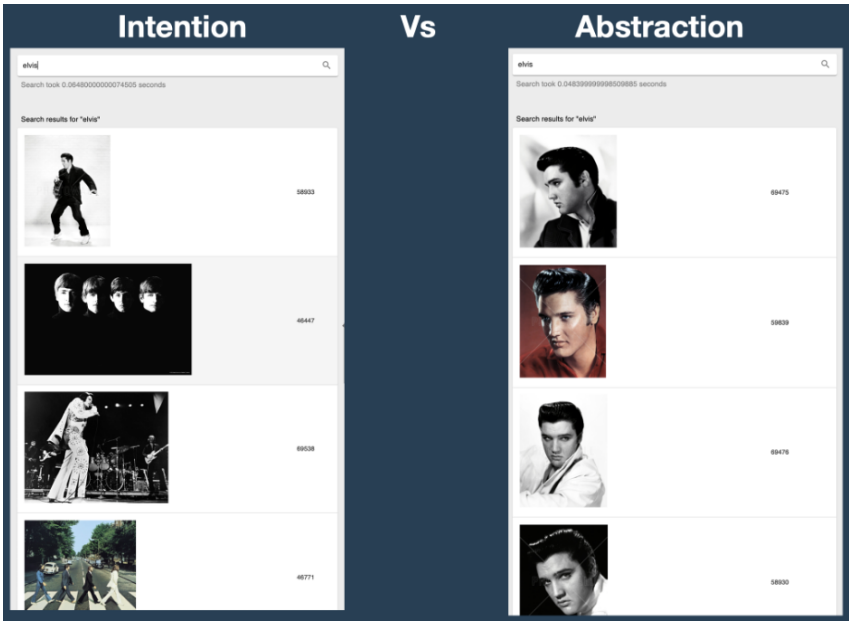

Figure 9: Intent (left) rather than abstraction (right) performs better.

Similarly, in Figure 10, we show how a generic LLM/vision model recommends reasonable old romantic movie motifs under a Clint Eastwood motif. However, this is completely missing the behavior that users exhibit, as they are in fact looking for masculine movie stars from the same era. Both are reasonable, but only one understands user behavior and drives sales.

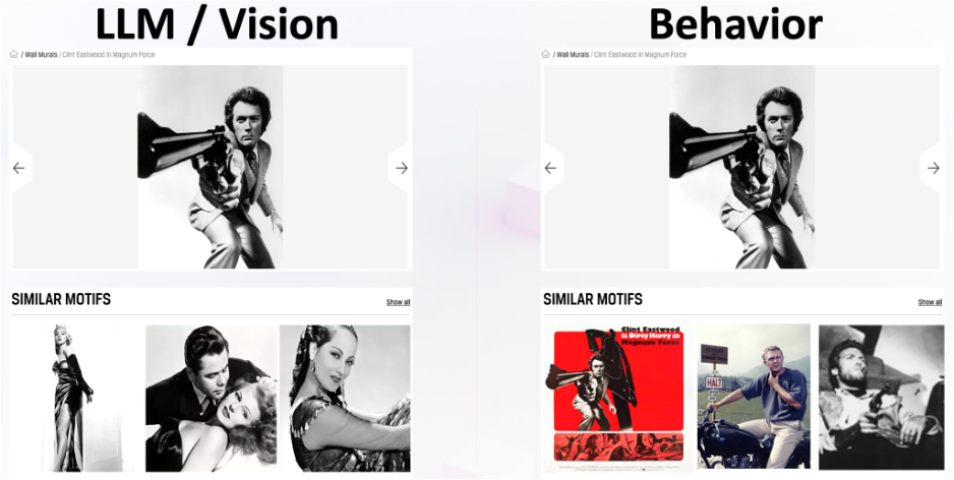

Figure 10: "Reasonable" non-contextual LLM/Vision model recommends black a white posters from romantic movies. BehaviorGPT that understands behavior, recommends the masculine movie stars from the same era that users are actually looking for.

We also demonstrate that we are maintaining very flexible abilities while we incorporate the contextualizer Transformer. In Figure 11, we show that we can search for many nuances about "Hepburn" and it is able to handle it very nicely, e.g., "Hepburn lying," "Hepburn smiling," etc. We credit our regularization technique for some of this graceful contextualization.

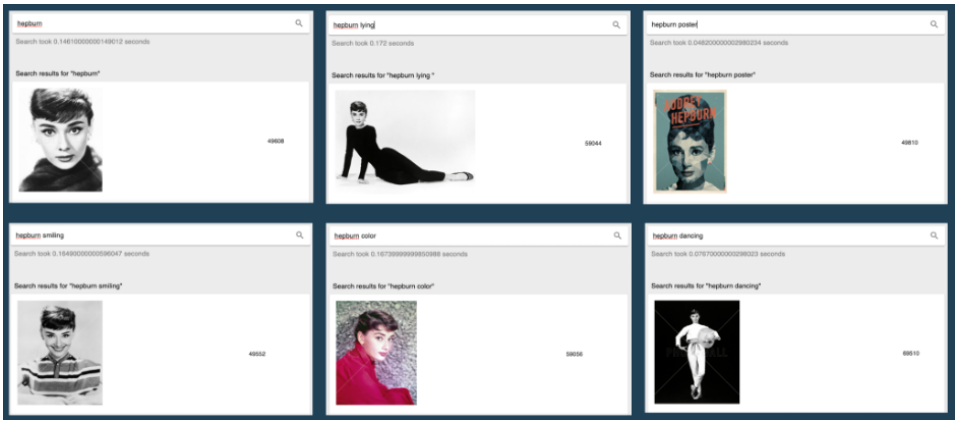

Figure 11: Nuanced and non-forgetting. BehaviorGPT learns to contextualize while maintaining very flexible abilities.

In Figure 12, we show expert labeling versus BehaviorGPT. Experts miss some of the most important keywords that users associate with these motifs, like “sunset” on top and “temple” and “Japanese” on bottom. It’s important to let users do the work for you and define the meaning of language.

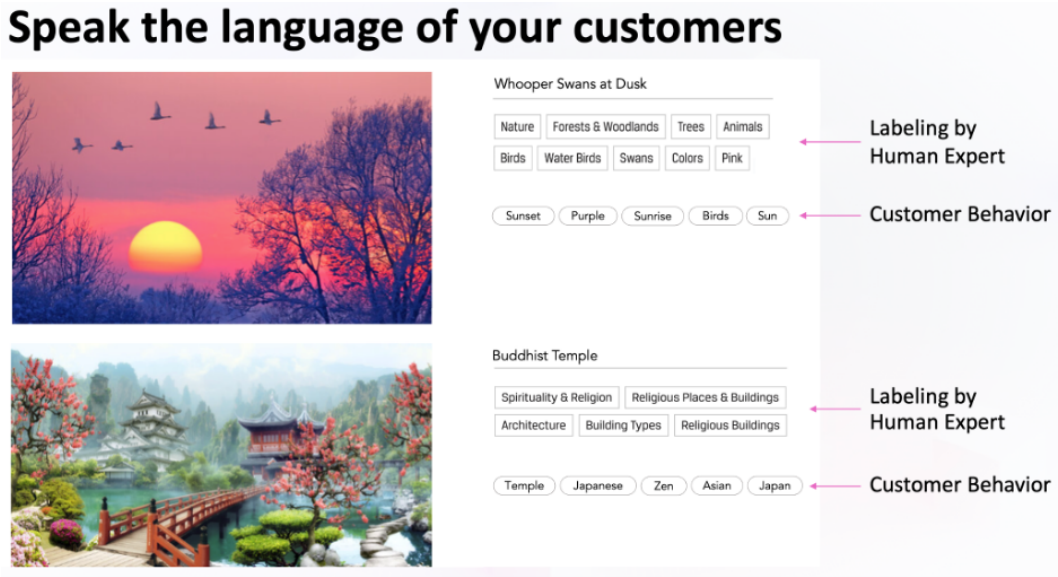

Figure 12: Expert labeling versus BehaviorGPT. Experts miss some of the most important keywords that users associate with these motifs, like “sunset” on top and "temple” and "Japanese” on bottom. It’s important to let users do the work for you and define the meaning of language.

As the famous saying goes, a picture says 1000 words, but still a lot of time we need to associate a list of keywords for an image. In Figure 13, we show BehaviorGPT can generate those words and rank them, so you know which ones are most important to users.

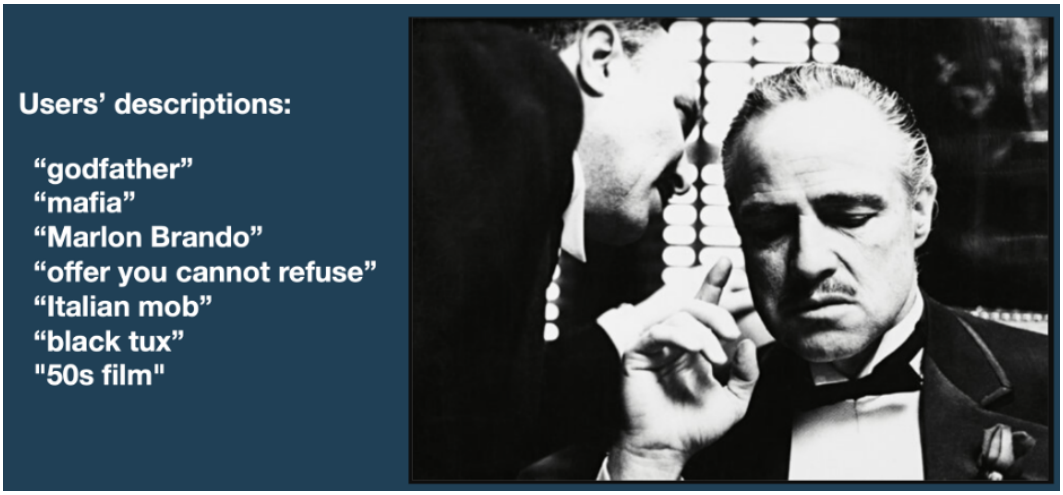

Figure 13: BehaviorGPT generated descriptions from user behavior on a motif named "Marlon Brando in the Godfather". It captures user behavior and more nuances.

Indeed, allowing users to explore visual art and aesthetics through language is important, as it is a natural way for us to search and explore categories. In Figure 14, we show how BehaviorGPT can suggest refined search terms based on a search that is optimized for user behavior. This can also be used to create and name categories, and it worked across around 50 languages and markets.

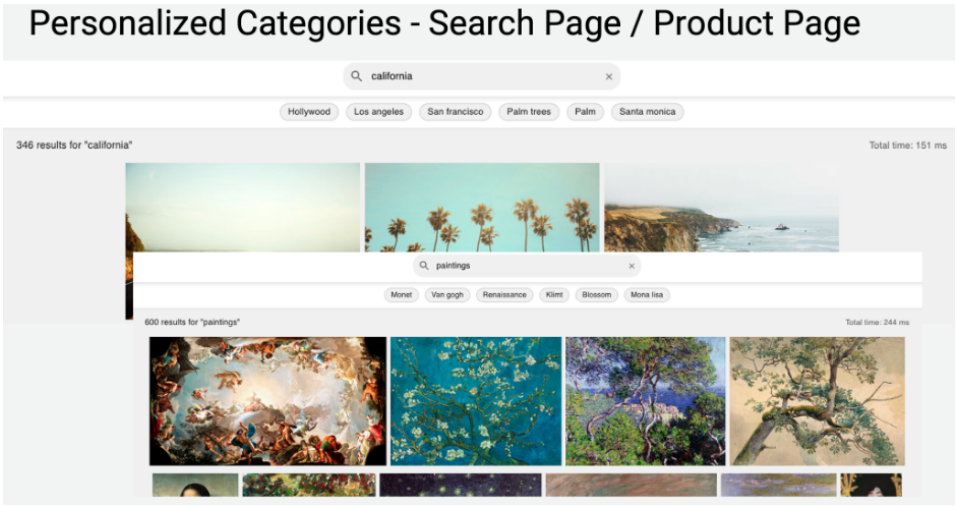

Figure 14: Examples of how BehaviorGPT understands the language that users use for visual art and aesthetics. It can use it to suggest refined search terms and for categorizing assortment.

We were also impressed how well the model learned to generalize. We could do cross-selling across companies, and BehaviorGPT was able to recommend relevant contextualized products out-of-domain. See Figure 15 for examples.

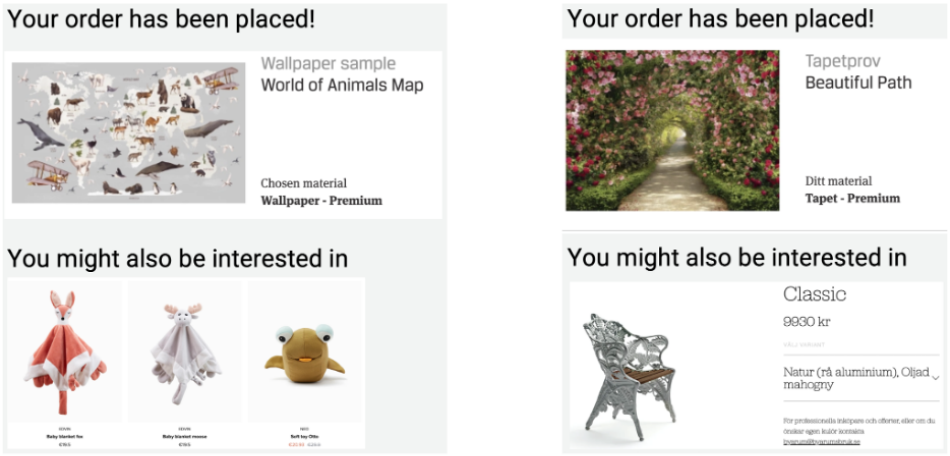

Figure 15: Generalization to out-of-domain motifs (products). From art and aesthetics, BehaviorGPT understands general aesthetic preferences like toys and furniture.

6.3 Personalized image generation

We can use BehaviorGPT for more generative abilities—yes, even to generate visual art and motifs themselves. The unique aspect is that we can generate images contextually based on any behavior. This can be some behavior + a text query, or simply just some previous behavior. The images will be tailored and optimized for the specific user. See Figure 16 for a diagram and a real generated example. We appended a Stable Diffusion model after BehaviorGPT and fine-tuned them together to generate the next motif given a history of behavior. We could condition the generation to be a wallpaper or canvas. Since attribution to designers is important to be able to share revenue, we could also ask BehaviorGPT what designer's art most contributed to this generation.

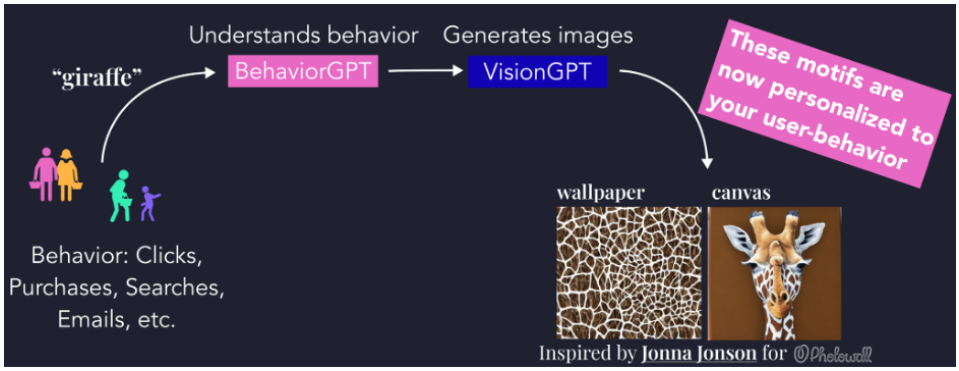

Figure 16: We appended a Stable Diffusion model after BehaviorGPT and fine-tuned them together to generate the next motif given a history of behavior. We could condition the generation to be a wallpaper or canvas. Since attribution to designers is important to be able to share revenue, we could also ask BehaviorGPT what designer's art most contributed to this generation.

In Figure 17, you can see when we try to use Stable Diffusion (VisionGPT) to generate similar variants non-contextually, versus when it has been fine-tuned together with BehaviorGPT to give much more relevant and contextual/personalized results.

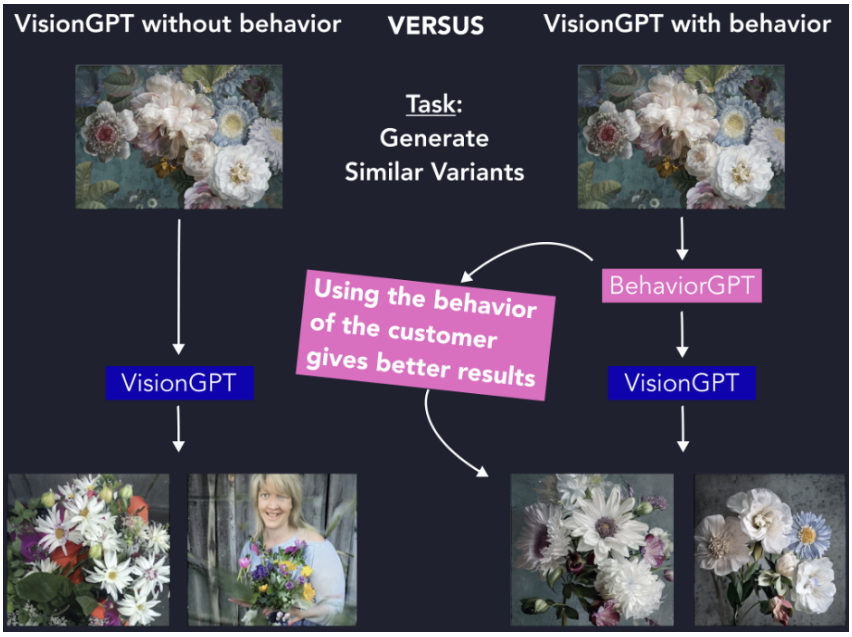

Figure 17: Stable Diffusion (VisionGPT) to generate similar variants non-contextually, versus when it has been fine-tuned together with BehaviorGPT to give much more relevant and contextual/personalized results.

6.4 Business intelligence

Now, BehaviorGPT is not only consumer-facing. The behavior of users tells you a lot of business intelligence, as it models your most important asset: your customers and your products.

In Figure 18, we plot a UMAP of user embeddings and color it by country. It shows interesting patterns; it mirrors geographical proximity to some extent, but very much cultural (or behavioral) proximity. Austria has a part in the middle and on each side of Germany. The USA is in the middle and shows more diversity in terms of intersecting with other countries and ultimately merging with the UK.

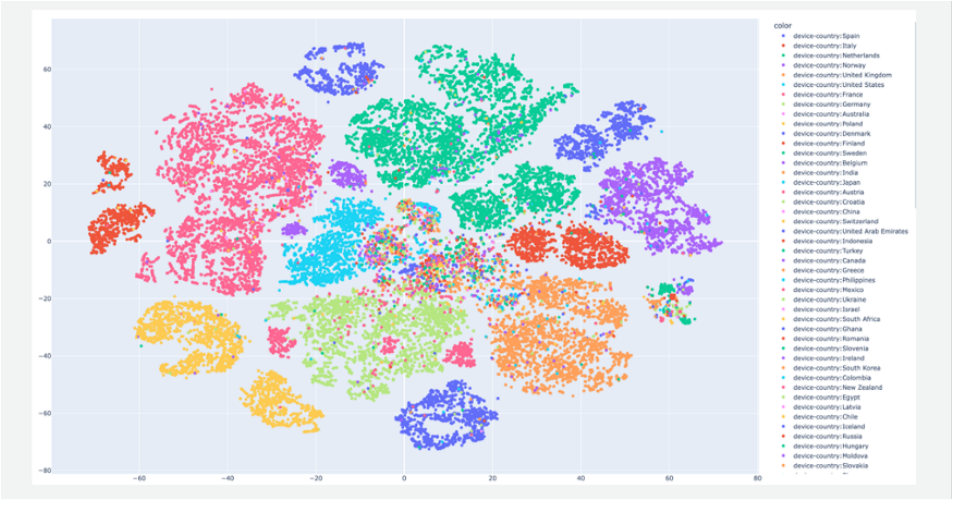

Figure 18: UMAP plot of user-embeddings colored by country.

In Figure 19, we show city embeddings as a UMAP plot. It was interesting to observe that the major cities within a country were clustered more closely together than geographical proximity would

suggest—indicating that urban vs. countryside has a bigger impact on behavior within a country than geography.

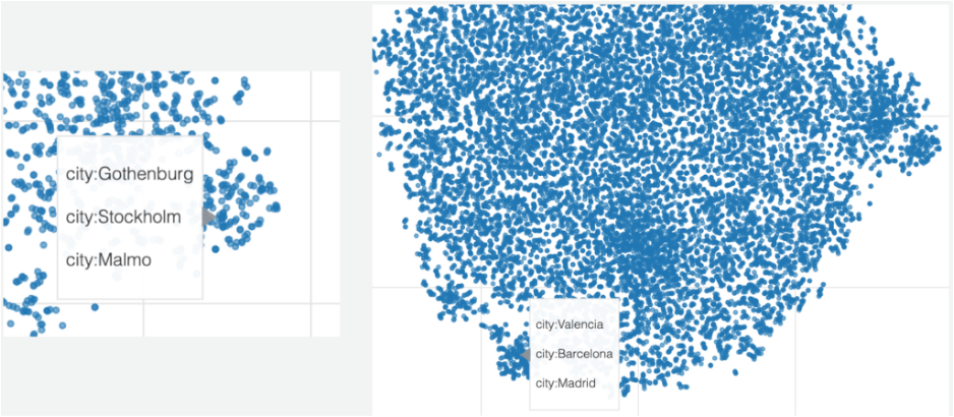

Figure 19: UMAP plot of embeddings of city domains. Big cities are mapped more closely than geographical proximity.

7 Ethical Considerations in Behavioral Aesthetics AI

As AI art markets grow to $50B by 2030 [5], ethical deployment is paramount. BehaviorGPT mitigates bias by grounding in real user behaviors across 50 markets, but we advocate for ongoing audits and fair use policies, ensuring AI augments—rather than displaces—human creativity. Our work on attribution of AI-generated content to designers is part of such work.

8 Conclusions

In summary, this work demonstrates the power of a behaviorally grounded foundation model for visual art and aesthetics, developed primarily between 2020 and 2023. By training a 0.5B-parameter Transformer on 215B actions derived from human interactions with art, we have established semantic understanding through behavioral co-occurrences, effectively decoding the "language" of aesthetics from action sequences rather than pixel patterns or static samples. In practical e-commerce deployment, this approach surpassed established specialized tools, delivering uplifts of +16% in search conversion, +24% in recommendations, +11% in dynamic categorization, and +14% in SEO and assortment optimization—all achieved with 12x fewer resources. Our qualitative evaluations further highlight the model's role in enhancing art exploration via natural language interfaces and personalized motif generation, while annual retraining revealed 2.5% compounding gains, underscoring the untapped "time-value of AI" in proprietary data ecosystems.

Additionally, integrating BehaviorGPT with image generation models enables context-aware creation with built-in designer attribution, promoting ethical practices. Ultimately, true behavioral intelligence in visual art emerges not from elicited textual descriptions but from observing unconscious actions—revealing deeper insights into user intent and opening avenues for future AI-driven creative augmentation by looking at what people do rather than what they say.

Citation

```
@article{unbox2025behaviorgptvisualart,
  author  = {Rickard Br{\"u}el Gabrielsson and Vasudev Gupta and
     others},
  title   = {BehaviorGPT for Visual Art: A Foundation Model for
     Aesthetics},
  journal = {Unbox AI Blog},
  year    = {2025},
  month   = jul,
```

```
  url      = {https://research.unboxai.com/behaviorgpt-visual-art-and-
      aesthetics.html}
}
```