Multimodal Representation Learning for User Qualia

ANN HE, Numinous Labs, Inc, USA

Modality gap in representation learning is a well-studied problem. While many have definitions of it in measurable benchmarks, most conceptualizations and solutions of the problem focus on style and lower-order concrete semantics. We elucidate a new perspective and class of problems in the space of multimodal representation learning, especially as it pertains to personalization, provide a proof of concept of finetuning a representation space for this problem, and discuss its applications in various generative AI pipelines.

CCS Concepts: • **Human-centered computing**; • **Computing methodologies** → **Artificial intelligence**;

Additional Key Words and Phrases: artificial intelligence, representation learning, media, philosophy

ACM Reference Format:
Ann He. 2025. Multimodal Representation Learning for User Qualia. 1, 1 (November 2025), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Qualia refers to the subjective, qualitative, and felt experiences of an individual's conscious experience. Examples include the feeling of pain, the taste of coffee, or the color red as a you, the individual, perceive it. Qualia is conditional on an individual's neural achitecture, so to speak, and the experiences they collect through their life, the particular environments they are embedded in (in the sense of other agents being part of an environment, in the sense of the Sapir Whorf Hypothesis [Whorf 1956], in the sense of Wittgenstein [Wittgenstein 1953]), and so on. (The Sapir-Whorf hypothesis suggests that language influences thought. This connects to Wittgenstein's concept of language games, where meaning emerges from use within specific contexts.)

In culture (so qualia of a collective group of individuals), such as literature and art, qualia can be described formally as synaesthesia (sensor crossover between modalities), aesthetic affinity (a form of emotional kinship), or semiotics (shared symbolic languages). Other informal words for this might be resonance, evocation, zeitgeist convergence (shared cultural moment expression).

2 Background

The concept of a modality gap was first introduced in Mind the Gap (Liang et al, 2022) [Liang et al. 2022] which posits that geometric inductive bias introduced in multimodal embeddings in which unimodal domains are tokenized and embedded separately creates a modality gap on image and image caption distributions.

Author's Contact Information: Ann He, annxhe@gmail.com, Numinous Labs, Inc, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM XXXX-XXXX/2025/11-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

A searchable continuous latent space which solves the modality gap lends itself to a multimodal embedding as well as a latent space for user-conditional multimodal generation. We believe that this is an approximation to understanding the phenomenal binding problem [Pearce 2012], which is about how objects, background objects, as well as abstract and affective features are integrated into a unified experience for an individual.

A motivating application of customizing a CLIP-like latent space is its use in custom text-conditional diffusion pipelines [Ramesh et al. 2022]. A custom latent space approach could be complementary to fine-tuning diffusion weights [Ruiz et al. 2022], which focuses more on direct style-transfer like results rather than semantic understanding.

Thus far, the representational learning research community has focused on multimodal distributions of (text, image) pairs which are relatively straightforward in their translation. For example, The Mind the Gap paper evaluates models for their geometric gap on the COCO dataset, [Lin et al. 2014] which contains photos of generic objects (in the same sense that [Ramesh et al. 2021] pre-trained DALL-E1 on image, text pairs for which the text appeared in Wikipedia > 100 times), Voyage, [Voyage AI 2024], evaluates mixed modality search on the distribution (text, image of text), i.e. that the string "Hello world" retrieves an image of "Hello world" rather than string such as "Cat."

While these evaluations form a baseline of multimodal representation gap, they still represent relatively simple cross domain transformations. For example, the transformation from image to image caption focuses on the object level, which most if not almost all observers of the image would agree on. And the transformation from text to image of text is as simple as save a pdf with "text" in it. In some sense, it means that these joint distributions have higher mutual information and are easier to learn than an individual's sensory space.

An individual's sensory space, on the other hand, is shaped by their histories, experiences, unique biology. Digitally, it is traceable, for example, through a user's hypertextual [Liu and Almeda 2025] space, intentional navigation through the web, manual linking, et cetera. When we learn a custom user adapter downstream of a pretrained baseline multimodal embeddings model, we are in, some sense, learning this transformation.

3 Dataset

The dataset we explore with is a second-degree scrape of a test user's text-based and image-based semantic space. This is chosen as a joint distribution which is representative of the test user's qualia (perhaps, in a cultural subspace).

We choose to join Substack and Pinterest space as they represent intentional content discovery platforms in which users explore an inner space such as aesthetics for home design, visualizing futures, or introspective writing with emotional qualities. In particular, these

are two domains in which we expect a high frequency of cross-artefact association driven by the user's intuitive style, or rhizomatic [Wikipedia contributors 2025] thinking.

In particular, this user-qualia-representative dataset is one in which various forms of a modality gap appear with embedding spaces such as voyage-multimodal-3.

3.1 Initial Evaluations on Raw Text and Image Features

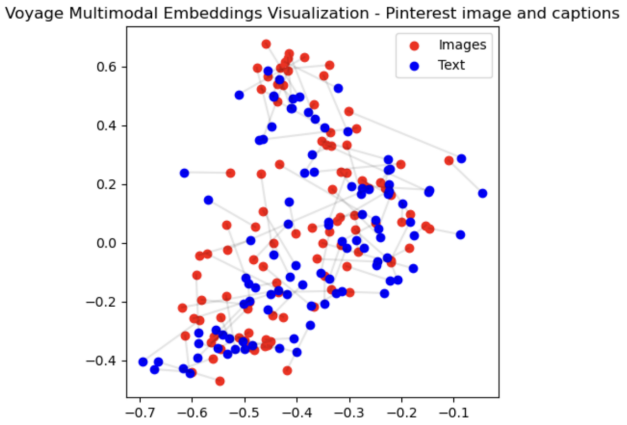


Fig. 1. Initial SVD on Pinterest images and captions

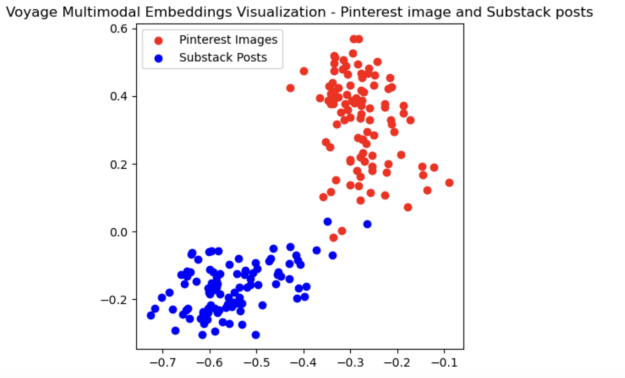


Fig. 2. Initial SVD on raw image and text features

The first evaluation above shows a 0.79 recall rate on Pinterest image and associated ground truth captions, which shows a good image-to-image-caption generalization ability of voyage-multimodal-3. The second evaluation above shows a clear modality gap on Substack posts, which are out of distribution of the (image, image-caption) distribution. In the raw feature space, text queries retrieve images of text regardless of the semantic content, rather than any other type of image artifact, thus failing the text to pdf of text evaluation in [Voyage AI 2024].

4 Methodology

4.1 Soft Labeling

Given the initial analysis of embedding raw text features, we choose to work, in this paper, with AI annotations of raw text and image features, so that the cultural knowledge of foundation models serves as a form of supervision signal. In future work, a finetuned model based on an individual could provide more attuned personalization. We provide, in the Appendix A.1, the prompts used.

The annotation allows us to project the raw image and text features to roughly the same space, when embedded with `voyage-multimodal-3` or `nomic-text-v1`, meaning the new feature space has good semantic overlap.

From this annotation semantic space, we run UMAP [McInnes et al. 2018] on the text and image domains individually, then HDBSCAN [Malzer and Baum 2019] to produce clusters. We then compute cluster assignments across domains via centroid similarity and filter to a 1-1 mapping via the Hungarian Algorithm [CP-Algorithms 2023]. Figure 3. Shows a visualization of the clustered space.

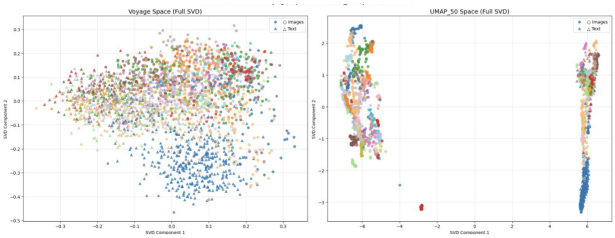


Fig. 3. Clustering with cluster labels

In Appendix A.2 we provide samples of semantic cross-modal clusters produced by this method.

4.2 Finetuning

LoRA [Hu et al. 2021] introduced the idea of sparse fine-tuning the through the addition of low-rank matrices. Inspired, we learn a linear and a two-layer adapter, which is applied to the output of `voyage-multimodal-3` to associate user qualia across text and image space.

The soft cluster labels provide positive-negative labels for fine-tuning. The initial loss function we use is the InfoNCE loss [van den Oord et al. 2018], written here for concreteness.

$$\begin{aligned}\text{InfoNCE}(q, X) = &- \log\left(\sum_{\text{pos}} \exp\left(\frac{\text{sim}(X_q, X_{\text{pos}})}{\tau}\right)\right) \\ &+ \log\left(\sum_{\text{neg}} \exp\left(\frac{\text{sim}(X_q, X_{\text{neg}})}{\tau}\right)\right) \qquad (1)\end{aligned}$$

Where q is an anchor point. For a given batch, we cycle through all points so each point is the anchor once in the computation of a per-batch loss. In the notation X_i represents a row-vector of the dataset. The loss function penalizes high similarity between the anchor and negative points while rewarding high similarity between the anchor and positive points.

5 Results

We measure few of initial metrics: cross-modal retrieval and cluster coherence (the ratio of retrieved artefacts which are of the same cluster), uniformity [Wang and Isola 2020], and create a query-histogram visualization for the effect of temperature on the fine-tuned distribution. The metrics are computed held-out clusters after the Hungarian Algorithm is applied to filter the cluster matching. We chose to hold out novel clusters rather than just data points as it is a more difficult generalization measurement. We note that the coherence jump from 1-layer to 2-layer is significant, as the Universal Approximation Theorem [Hornik et al. 1989] would suggest, and demonstrate on these two architectures as a proof of concept. The uniformity score is measured as it is associated with downstream retrieval performance and the author's choice to measure it was due to initial observations on the uniformity of the query histograms.

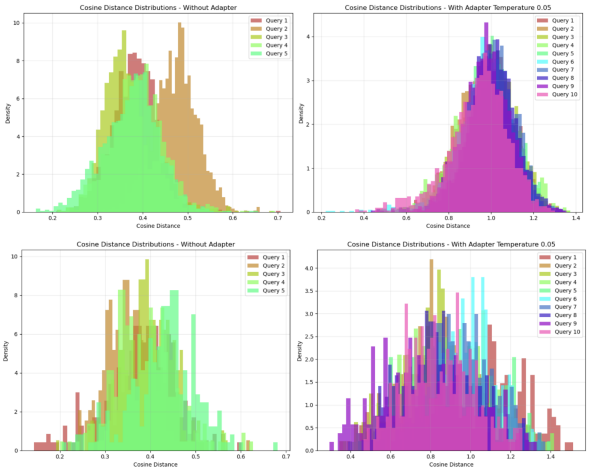


Fig. 4. Query histogram of random queries, top: one layer adapter, bottom: two layer adapter

Embedding Space	Uniformity
voyage-multimodal-3-text-annotations	-1.563
1-layer-adapted-temp-0.05	-2.968
1-layer-adapted-temp-0.1	-3.712
1-layer-adapted-temp-0.2	-3.760

Table 1. Uniformity scores for different embedding spaces.

Initially we had a coherence result of about 0.105 ± 0.101 on both train and validation because we initially computed UMAP clustering separately on train/validation splits, but this creates incomparable cluster assignments. We corrected this by computing clusters on the full dataset before splitting matched pairs. Analyzing the train set coherence helped us diagnose the methodological issue.

Since the initial results show a tradeoff between modality balance and coherence, we further experiment with adding a modality balance term to the noise contrastive loss, so the objective is:

Temperature	Avg (text/image) Ratio
Without Adapter	6.852
1 Layer 0.05	2.350
1 Layer 0.1	2.668
1 Layer 0.2	4.010
2 Layer 0.05	5.591
2 Layer 0.1	6.416
2 Layer 0.2	7.580

Table 2. Average text/image token ratio across temperature and adapter configurations.

Temperature	Coherence
Without Adapter	0.36 ± 0.190
1 Layer 0.05	0.408 ± 0.143
1 Layer 0.1	0.395 ± 0.160
1 Layer 0.2	0.355 ± 0.140
2 Layer 0.05	0.815 ± 0.305
2 Layer 0.1	0.887 ± 0.201
2 Layer 0.2	0.690 ± 0.293

Table 3. Coherence scores (± std. dev.) for different temperature and adapter settings.

$$\text{InfoNCE}(q, X) = -\log\left(\sum_{\text{pos}} \exp\left(\frac{\text{sim}(X_q, X_{\text{pos}})}{\tau}\right)\right) + \log\left(\sum_{\text{neg}} \exp\left(\frac{\text{sim}(X_q, X_{\text{neg}})}{\tau}\right)\right) \tag{2}$$

$$+ \frac{\text{retrieved of same modality as q}}{\text{retrieved of different modality as q}} \tag{3}$$

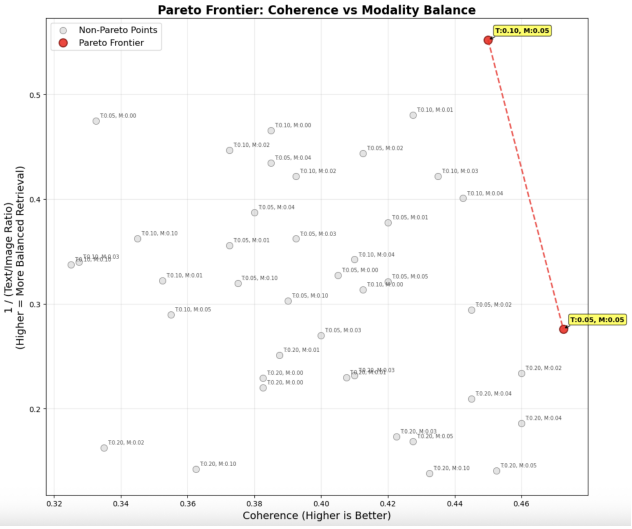


Fig. 5. Pareto Frontier between coherence and modality balance

We plot the results of optimizing with the new objective and highlight the temperature and modality balance weight objectives which have the best balance on these two metrics for the two-layer mlp. All other hyperpameters are kept constant.

We note that coherence is relatively low when the trade-off is made on this dataset with modality balance (compared to the 0.88 coherence observed), which could be an artifact of naively computing modality balance on in-batch elements without accounting for which cluster the examples belong to. We leave further experimentation with the contrastive loss, including adding terms for a generative captioning loss to future work, where we also modify the architecture which produces the multimodal representation space.

6 Acknowledgments

The author acknowledges Bushra Farooqui, Jennifer She, Jeremy Nixon, and Lily Nguyen for discussion which have lead to ideas in this paper.

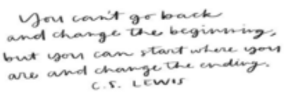
A Research Methods

A.1 Prompts for Annotation

```
ANNOTATION_PROMPT = """You are analyzing Substack posts.
     Provide both content summary and cultural analysis.

OBJECT LEVEL DESCRIPTION: Summarize the main topic, argument,
      or narrative in 5-6 sentences.

CULTURAL ANALYSIS:
1. AESTHETIC: [underlying aesthetic/intellectual sensibility
     ]
2. THEMES: [3-5 philosophical/cultural concepts]
3. MOOD: [emotional register/tone]
4. LIFESTYLE: [way of being this suggests]
5. MOMENT: [broader cultural conversation]

Format your response as:
SUMMARY: [content summary]
AESTHETIC: [aesthetic philosophy]
THEMES: [concept1, concept2, concept3, concept4, concept5]
MOOD: [emotional register]
LIFESTYLE: [lifestyle implications]
MOMENT: [cultural moment]

Focus on the deeper cultural intelligence - what cultural
     movements, aesthetic philosophies, or ways of being
     this writing represents or advocates for."""
```

A.2 Cluster Examples

1848: a black and white photo with a handwritten quo



1941: the words are written in cursive writing on a ink

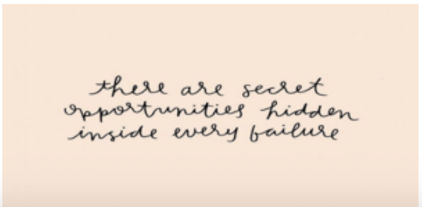


Fig. 6. Cluster A Image Exemplars

```
Text Cluster 27 Exemplars:
----------------------------------------
  1157: ruthreichl - The Less Ground We Have, the Better We Can Cultivate It


  246: www.frederikjournals.com - ⁄Writing Walks and the Awe of Spring
       One of the fun things about meeting my subscribers has been the uncertainty. I
never know where the conversation will go or how it will connect with a topic I'm writ
ing about. Recently, I met someone who has been on an insane walking streak: 2-5 miles
a day since 2011 ! Every. Single. Day! We talked...

  251: www.frederikjournals.com - The weight and the mask.
       I ripped up a book. I ripped up a book and now I can't sleep. Something came ov
er me, burst out of me. A wave of anger. A tide, ancient, boiling, bubbling. I took th
e book and slammed it against the wall. I hammered it against the floor until it fell
apart, until pages flew. What the fuck. I love bo...
```

Fig. 7. Cluster A Text Exemplars

A References

References

CP-Algorithms. 2023. Hungarian Algorithm. Retrieved September 4, 2025, from https://cp-algorithms.com/graph/hungarian-algorithm.html.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 5 (1989), 359–366.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685* (2021).

Wei Liang, Yujia Zhang, Yewon Kwon, Serena Yeung, and James Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. *arXiv preprint arXiv:1405.0312* (2014).

Shirley Liu and Santiago G. Almeda. 2025. Agency Among Agents: Designing with Hypertextual Friction in the Algorithmic Web. *arXiv preprint arXiv:2507.23585* (2025).

Christoph Malzer and Michael Baum. 2019. A Hybrid Approach To Hierarchical Density-based Cluster Selection. *arXiv preprint arXiv:1911.02282* (2019).

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426* (2018).

David Pearce. 2012. Non-materialist physicalism: an experimentally testable conjecture. Available at https://www.physicalism.com/.

Fig. 8. Cluster B Image Exemplars

```
Text Cluster 3 Exemplars:
----------------------------------------
  1351: sarapetersen — The Secret Lives Of Women Who Snap
       In her new novel, Madwoman (which someone needs to adapt into a Netflix series
stat), Chelsea Bieker provides her reader with nearly everything a bookish soul could
want. Bitingly smart satirical analysis about wellness culture and the allure of the s
tatus grocery store? Check. Raw excavations into ...

   792: www.evilfemale.blog — Love In The Time Of True Crime Podcasts
       I thought I was going to die outside the public library. It was a shame—I've al
ways been fond of the library as a place of community and the Troy Public Library was
a particularly gorgeous one. I had called ahead that day asking if I could come take s
ome pictures for an art project. It was a concept...

  1441: teachrobotslove — A Lonely Corpse Waiting For Dawn
       NOTE: This is Part 2 of my 5-part Evil Series. If you like this kind of writin
g, I'd love to have your support via a paid subscription . I need to buy some new sock
s and a bottle of tequila. "People who cease to believe in God or goodness altogether
still believe in the devil... Evil is always possi...
```

Fig. 9. Cluster B Text Exemplars

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *arXiv preprint arXiv:2102.12092* (2021).

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *arXiv preprint arXiv:2208.12242* (2022).

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

Voyage AI. 2024. voyage-multimodal-3: all-in-one embedding model for interleaved text, images, and screenshots. https://www.voyageai.com/blog. Voyage AI Blog.

Tongzhou Wang and Phillip Isola. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. *arXiv preprint arXiv:2005.10242* (2020).

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* MIT Press.

Wikipedia contributors. 2025. Rhizomatic learning. Retrieved September 4, 2025, from https://en.wikipedia.org/wiki/Rhizomatic_learning.

Ludwig Wittgenstein. 1953. *Philosophical Investigations.* Blackwell. Original work published posthumously.