PERSISTENT REPRESENTATIONAL MISALIGNMENT INDUCES VISION-LANGUAGE MODEL FAILURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) fail at some tasks that are simple for humans—but why? We hypothesize that many such failures stem from VLMs' difficulties translating between vision and language representations—a misalignment the Platonic Representation Hypothesis predicts should disappear with scale. This paper is composed of two parts: in the first, we show that existing vision and language representations are linearly unalignable. This misalignment is an inherent property of the data and hence expected to persist regardless of either scale or the alignment's complexity. In the second, we demonstrate how this misalignment both leads to a predictably induced generalization failure and correlates with existing failure modes, where more representationally aligned VLM backbones accordingly exhibit fewer failures. Overall, these results imply representations will not become aligned simply given sufficient scale, providing evidence against a strong version of the Platonic Representation Hypothesis.

1 INTRODUCTION

Although vision and language models are considered superhuman on many tasks (He et al., 2015; Achiam et al., 2023), their multimodal counterparts exhibit some very simple failures. Vision-language models (VLMs) struggle identifying visual analogies (Yiu et al., 2024), matching semantically different but syntactically similar captions to images (Thrush et al., 2022), identifying simple geometric properties like whether two circles overlap (Rahmanzadehgervi et al., 2024), and counting atypical objects, e.g. the number of legs on a three-legged chicken (Vo et al., 2025). Humans outperform VLMs on all of these tasks.

What causes these failures and why they persist is generally unclear, with several linked observations. Vo et al. (2025) observe that VLMs can rely on their language component's world knowledge ("chickens typically have two legs") more than on what their vision backbones see, but do not discuss why this occurs. Testing why VLMs struggle with fine visuolingual retrieval, Diwan et al. (2022) speculate that it is due to misalignment between the visual and text representations.

A lack of representational alignment is surprising given some common beliefs in the field. The Platonic Representation Hypothesis states that sufficiently well trained models should converge to equivalent representations (Huh et al., 2024), with many works presuming or relying on such convergence (Tjandrasuwita et al., 2025; Zhu et al., 2025; Jha et al., 2025; Gupta et al., 2025; Wang et al., 2025b; Bahng et al., 2025; He et al., 2025).

In this paper, we call the Platonic Representation Hypothesis into question and show implications thereof on existing models. To test whether vision and language representations converge, we develop a simple theoretical framework for when sets of representations are alignable (Section 3). We find that existing representations are unalignable due to underlying differences in the data, meaning alignment will likely not naturally emerge with scale. This misalignment predictably leads to an induced generalization failure in vision-language models, where a VLM fails to recognize a concept its vision and language backbones know if it does not appear in its paired finetuning data (Section 4). Moreover, representational alignment correlates with performance on the failure modes, where more aligned VLM backbones lead to VLMs which exhibit fewer failures (Section 5).

Altogether, these results support the conjecture that **a diverse set of VLM failures likely stem from representational misalignment and will not be resolved solely by increasing model scale**. This conjecture provides a joint cause for seemingly disjoint failure modes while explaining why the failures persist even in very capable models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

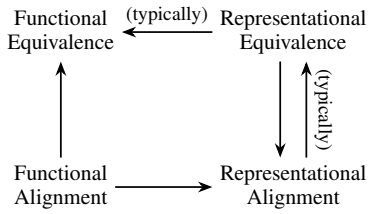
2 Relations Between Representations

In this section we discuss different ways in which sets of representations can be similar. While there are many existing notions (Sucholutsky et al., 2023), many are statements on whether the representations are *equivalent* or *alignable*. We now discuss what each of these mean and the relationships between different kinds of similarity, with additional related work in Appendix A.

There are two main ways in which different sets of representations can be equivalent. The first is *functional equivalence*, where two representations or models producing some representations are used for some task. For example, take two sets of word embeddings used for retrieval. In this case, the representations are equivalent if they perform comparably well and make the same mistakes, thereby being indistinguishable from a black-box perspective. A weaker but related form of equivalence is *representational equivalence* (Klabunde et al., 2025), where sets of representations are equivalent up to some transformation. Invariances to different transformations define different similarity kernels between representations, with most similarity kernels defining two representations as equivalent if they are the same up to some (constrained) linear transformation, perhaps with some additional normalization (Kornblith et al., 2019, section 4). Works like the Platonic Representation Hypothesis primarily discuss representational equivalence (Huh et al., 2024). Representational equivalence is likely, but not guaranteed, to result in functional equivalence – for the word embeddings, an orthogonal transformation over an entire space would not affect which words are retrieved as distances would be preserved, but more general transforms could lead to functional differences.

Representational equivalence is closely related to *representational alignment*. For typical generative VLMs, pretrained vision and language models are combined using an adapter that projects vision representations into the language model, thereby aligning the vision representations with the language token embeddings (Alayrac et al., 2022; Merullo et al., 2022; Liu et al., 2023). For retrieval, language and vision representations are aligned by being projected into a shared representation space (Radford et al., 2021). Representations that are representationally equivalent are easy to align, as a linear map is often sufficient to map between them. More broadly, representations can be *functionally alignable* if given some training they can be used together to generalize over some task, where generalization is important as an adapter may just memorize training data. If the adapter is a simple linear map, as it often is for generative VLMs (Liu et al., 2023; Zhu et al., 2023), functional alignment implies a degree of representational alignment, as the learned map aligns the representations. This is often used in model-stitching (Lenc & Vedaldi, 2015; Bansal et al., 2021; Csiszárik et al., 2021), where parts of one model are stitched to another using a trained linear layer, as a way to probe whether two models learn the same representations.[1]

In practice, both equivalence and alignment are not binary properties but fall on various spectrums and in some cases one can imply another. Typical relationships between equivalence and alignment are summarized in this diagram, where arrows represent "X implies Y" and "typically" means that counterexamples exist but are somewhat adversarial. Thus, two representations being representationally equivalent typically implies they are functionally equivalent as well.



This framework helps understand how different statements on representations relate to one another. For example, the Platonic Representation Hypothesis (Huh et al., 2024) asserts that representations in different modalities should be representationally equivalent and tests this by showing trends in representational alignment. Representational alignment typically implies representational equivalence, thereby supporting their claim. Throughout this paper we will show different forms of alignment or equivalence fail to hold, noting how one not holding rules out another, e.g. representations cannot be functionally aligned if they are functionally inequivalent.

3 Broken Conditions for Representational Alignment

In this section we derive the conditions for representations to be linearly alignable. Operationally, many alignment metrics like CKA quantify whether sets of representations are linearly alignable up to normalization, restricted to some subset of linear transformations (Kornblith et al., 2019). In Appendix B we formalize some of this intuition, showing how certain bounds on representational alignability accordingly bound some common representational similarity metrics.

[1] This reasoning has some known caveats; see Smith et al. (2025).

A useful way to test for representational alignability is by considering how concepts are related in different representations (Dhimoïla et al., 2026). For example, over a decade ago Mikolov et al. (2013b) showed that word embeddings trained on different languages have similar relationships between the same sets of words, so a linear map is sufficient to translate between representations in different languages. In this case each word embedding defines how a concept, namely its word, is represented.

One very simple kind of relationship between concepts is analogies. More formally, let there be some analogy where A is to B as C is to D, e.g. king is to queen as man is to woman. The vision representation of a concept like "king" can be found by averaging the representations of many images of kings. All presented theorems are based on simple known lemmas in linear algebra, with proofs delegated to Appendix C. If v_X denotes some model's representation of concept X then we can formally define an analogy as follows.

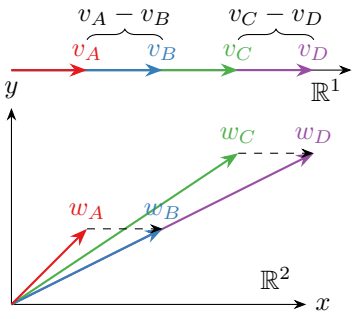
Definition 3.1 (Concepts and analogies in representations)**.** The vector representing some concept X is denoted as v_X. The analogy A is to B as C is to D (denoted $A : B :: C : D$) is said to exist for a model's representations of these concepts if $v_A - v_B = v_C - v_D$. The tuple representing an analogy is denoted as $v^{A:B::C:D} := (v_A, v_B, v_C, v_D)$. We assume no concept is represented by the zero vector.

In practice, the analogy would only be approximate, so $v_A - v_B \approx v_C - v_D$. Throughout this section, take V, W to be two finite-dimensional inner product spaces, e.g. $\mathbb{R}^n$ and $\mathbb{R}^m$. One can intuitively think of them as the representations of two different models, potentially from different modalities. For simplicity, all inner products are viewed as dot products, where $\langle v, w \rangle = v^T w$, but the theorems equally hold regardless of the kind of inner product.

If there are enough corresponding analogies between two models then one might expect a linear map between their representations to exist, as the analogies enforce a linear structure. In practice, this is not the case.

Theorem 3.2 (Analogies are insufficient for a linear map to exist)**.** *Let there be two sets of the same analogies in V and W, $\{v^{A_i:B_i::C_i:D_i}\}$ and $\{w^{A_i:B_i::C_i:D_i}\}$. There exist sets of analogies such that there is no corresponding linear map $T : V \to W$ such that $\forall i, X : T(v_{X_i}) = w_{X_i}$.*

The following example is effectively a proof of this theorem.



Example 3.3 (Different information can lead to inconsistent mappings)**.** Observe the following analogies in $\mathbb{R}^1$ and $\mathbb{R}^2$, $v_A = [1], v_B = [2], v_C = [3], v_D = [4]$ and $w_A = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], w_B = \left[\begin{smallmatrix}2\\1\end{smallmatrix}\right], w_C = \left[\begin{smallmatrix}3\\2\end{smallmatrix}\right], w_D = \left[\begin{smallmatrix}4\\2\end{smallmatrix}\right]$. Clearly, there is no linear transformation $T(v_X) = w_X$ as $4v_A = v_D$ but $4w_A \neq w_D$. The diagram to the right illustrates this visually.

Example 3.3 can be equivalently interpreted as one representation having information that is missing in the other. Intuitively, the second index in the w vectors could be some information that is irrelevant for the analogy but consistently exists, like a typical syntax in a sentence or different lighting in an image.

A stronger requirement, which does result in linear mappings, is that the different representations have the same similarity structure, akin to the case of Mikolov et al. (2013b) in machine translation (Smith et al., 2017). For representations with unit norm this means that the cosine similarity between concepts A and B is the same in both modalities, so $v_A^T v_B = w_A^T w_B$. For example, this condition implies that the cosine similarity between an image of a cat and an image of a dog is close to the cosine similarity between the words "cat" and "dog".

Definition 3.4 (Similarity structures)**.** Given a set of concept representations $\{v_{X_i}\}_{i=1}^N$, the Gram matrix $G_{ij} := v_{X_i}^T v_{X_j}$ is said to define the representations' similarity structure. Two representations are said to have the same similarity structure if their Gram matrices are the same, so $\forall i, j : v_{X_i}^T v_{X_j} = w_{X_i}^T w_{X_j}$.

Interestingly, having the same similarity structure implies that under some similarity metrics, like CKA, these representations are equivalent.

Theorem 3.5 (Same similarities enable linear mappings)**.** *Let V and W have corresponding sets of concept representations, $\{v_{X_i}\}_{i=1}^N, \{w_{X_i}\}_{i=1}^N$. If they have the same similarity structure then there exists a linear map $T : V \to W$ such that $\forall i : T(v_{X_i}) = w_{X_i}$.*

This theorem is stronger than required for a linear map to exist, as the resulting T is an isometry, a map that preserves distances. Moreover, assuming the spaces have the same dimension, such a map would have an inverse, so it loses no information as it allows mapping back and forth between them. In the single modality setting Smith et al. (2017) construct such a map for translating between sets of word embeddings, showing that the different embeddings indeed have a similar similarity structure. Representations having different similarity structures, however, restricts which transformations can map between them.

Corollary 3.6 (Different similarity structures rule out isometric mappings)**.** *Assume V, W have corresponding sets of concept representations but different similarity structures. Thus, there cannot exist an isometric linear map $T : V \to W$ such that $\forall i : T(v_{X_i}) = w_{X_i}$.*

A weaker condition that results in generic linear maps is requiring a form of generalized analogies to exist in both modalities. For example, if $v_{\text{zebra}} = v_{\text{horse}} + v_{\text{stripes}}$ then $w_{\text{zebra}} = w_{\text{horse}} + w_{\text{stripes}}$ as well.

Theorem 3.7 (Generalized analogies enable linear mappings)**.** *Let V and W have corresponding sets of concept representations, $\{v_{X_i}\}_{i=1}^N$ and $\{w_{X_i}\}_{i=1}^N$. A linear map T such that $\forall i : T(v_{X_i}) = w_{X_i}$ exists if and only if for all sets of scalars $a_i \in \mathbb{R}$ if $\sum_i a_i v_{X_i} = 0$ then $\sum_i a_i w_{X_i} = 0$.*

Note that any generalized analogy can be written as $\sum_i a_i v_{X_i} = 0$, e.g. in the case of $v_{\text{zebra}} = v_{\text{horse}} + v_{\text{stripes}}$ by moving v_{zebra} to the right hand side. A natural but important corollary of this theorem is that regular analogies not holding results in generalized analogies not holding as well.

Corollary 3.8 (Generalized analogies result in and require regular analogies)**.** *If the conditions of Theorem 3.7 hold then any regular analogy must exist in both V and W as well, so if $v^{A:B::C:D}$ is an analogy then so is $w^{A:B::C:D}$. Moreover, any analogy existing only in V but not in W, so $v_A - v_B = v_C - v_D$ but $w_A - w_B \neq w_C - w_D$, violates the conditions of Theorem 3.7.*

In practice, text and vision have neither the same similarity structure nor the same analogies, respectively violating the conditions of both Theorem 3.5 and Theorem 3.7. Note that, as Theorem 3.7 is both a necessary and sufficient condition, this implies there is no concept-preserving linear map between them. We first show that text and vision representations have a different similarity structure and then that they have different analogies.

3.1 Language and Vision Have Different Similarity Structures

Similarity between datapoints is typically measured using representations from semantic similarity models. When these representations are normalized the dot product between them corresponds to a cosine similarity. For text and vision to have the same similarity structure (as per Definition 3.4), thus fulfilling the conditions of Theorem 3.5, two images with a given cosine similarity should have captions with approximately the same cosine similarity as well.

In practice, visual and textual descriptions of the same scene can be similar in one modality but very different in another due to the different modalities having different information contents. Figure 1 gives some examples of image-caption pairs where, due to each modality containing different information, this occurs. All image-caption pairs are from the COCO training set. Cosine similarities are computed using DINOv3 base for the images and Sentence Transformers' all-MiniLM-L12-v2 for the captions (Siméoni et al., 2025; Reimers & Gurevych, 2019). Caption representations are averages over all of an image's caption variants in the dataset.

Pairs like those in Figure 1 occur with a non-negligible frequency. Their prevalence can be quantified by finding pairs for which the cosine similarity between the images differs substantially from that between their captions, with the absolute difference in their cosine similarities being above some large threshold (e.g. 0.4). For about 2% of COCO image/caption pairs, their cosine similarity difference exceeds this threshold.

To quantitatively measure whether the two representations have the same similarity structure, we empirically compare their cosine similarities. Formally, for models



Figure 1: **Two scenes can be similar in one modality but different in another. (left)** shows two images which are visually similar, where a vision embedding model gives them a cosine similarity of 0.96. Their corresponding captions are more distinct, having a similarity of 0.44. **(right)** shows the opposite case, where the captions are similar but the images are distinct. The captions have a cosine similarity of 0.97 while the images' cosine similarity is -0.06. The text similarities are for the average embeddings over the five different captions given in COCO for each image. See Appendix D for more examples.

f_1, f_2 outputting normalized vectors $f_1(x), f_2(x)$, we define $\text{CosSimRMSE}(f_1, f_2) := \sqrt{\mathbb{E}[(f_1(x)^T f_1(y) - f_2(x)^T f_2(y))^2]}$, where the expectation is over the dataset. CosSimRMSE tests the extent to which two representations have the same similarity structure as per Definition 3.4. Although here both models get as input some x, in general one model could get an image while the other gets the corresponding set of captions. We measure $\text{CosSimRMSERatio}(f_1, f_2 | g_1, g_2) := \frac{\text{CosSimRMSE}(f_1, f_2)}{\text{CosSimRMSE}(g_1, g_2)}$, with g_1, g_2 representing two models in the same modality that form some lower bound. As Table 1 shows, vision and text models have an approximately $3\times$ larger gap between one another than models in the same modality. CLIP models, despite being purposely trained to align their representations, exhibit only a slightly better degree of alignment likely due to this difference being inherent in the data.

Table 1: Relative cosine similarity RMSEs for representations from different vision and language models. RMSEs across different modalities are typically at least $3\times$ higher than those within modalities. All values are relative to Text,Text', where $\text{CosSimRMSE}(\text{Text}, \text{Text'}) = 0.05$. RMSEs are calculated using image/caption pairs from the COCO dataset. Models are DINOv3-B and -S for Vis and Vis', all-MiniLM-L12-v2 and -L6-v2 for Text and Text', and the CLIP-B/16 vision/text encoders for CLIP Vis/Text.

Representations		Cosine sim. RMSE ratio
Text	Text'	1.0
Vis	Vis'	1.3
Vis	Text	3.3
Vis	CLIP Vis	10.9
Text	CLIP Text	11.2
CLIP Vis	CLIP Text	2.9

Altogether, this section's results show that Theorem 3.5's condition does not hold, as vision and language have very different similarity structures. We now show that Theorem 3.7 also cannot hold, showing that typical vision and language representations are linearly representationally unalignable.

3.2 Language and Vision Have Different Analogies

To see whether vision and language representations exhibit the same analogies we use the text analogy dataset of Mikolov et al. (2013a). These include analogies like king:queen::man:woman and Paris:France::London:UK. To compare lingual to visual analogies, we construct a new dataset called VisLingA (for **vis**ual **ling**ual **a**nalogies), containing images representing various concepts, see Appendix E for qualitative examples. VisLingA is constructed by downloading 5-10 images for each concept in the text analogies dataset. Analogies containing words with too few images and analogies

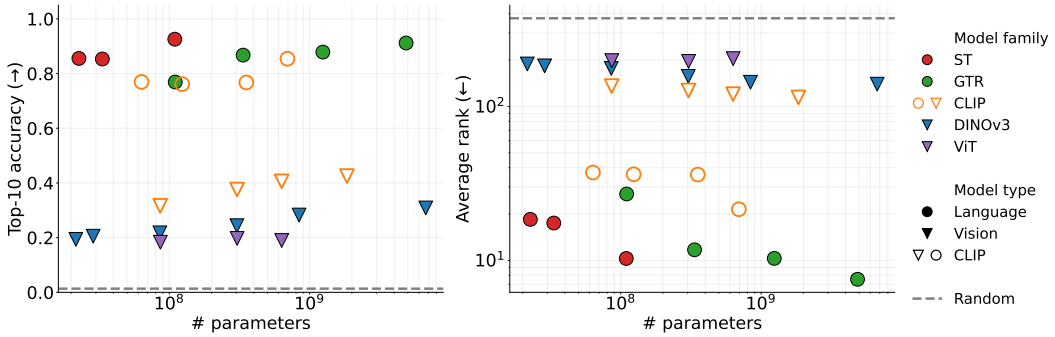


Figure 2: **Vision models do not exhibit the same analogies that language models do**. (left) top-10 analogy retrieval accuracies over VisLingA for several model families. The vision models (downwards triangles) do significantly worse than the text models (circles). (right) average retrieval rank. All vision models have a retrieval rank above 100, with a random baseline having a rank of 373.5. In both cases, the CLIP's text encoders are the worst models in the language family and the opposite for the vision CLIP models, likely due to the intermodal supervision pulling each CLIP encoder towards the other modality. ST=Sentence Transformer (Reimers & Gurevych, 2019), GTR is from Ni et al. (2022), CLIP is the LAION CLIP models (Cherti et al., 2023), ViT are ViT models trained on ImageNet 21k (Dosovitskiy et al., 2020).

in the plural, present-participle, and adjective-to-adverb categories are discarded as they tend to contain different words with mostly the same pictures, e.g. "car" and "cars". Images were downloaded from Wikimedia Commons and filtered to images with CC0 or CC-BY copyright licenses, keeping creator metadata. The final dataset has 15,455 analogies containing 746 distinct concepts.

An analogy $A : B :: C : D$ can be said to exist in a modality if the representation of $A - B + C$ is close to D. Retrieval is done over the embeddings for all the concepts in the dataset, so representations exhibit a set of analogies if they have a low average rank or high accuracy. Vision embeddings are found by averaging all of a concept's images' representations.

Results. Figure 2 shows that while most analogies exist in text, far fewer are present in vision. All text embedding models have top-10 analogy retrieval accuracies above 75% while all vision models get below 45%. This is likely due to the inherent difficulty of capturing abstract concepts or certain nouns in images, like countries and cities. Thus, as vision and language models exhibit different analogies, Theorem 3.7 cannot hold.

4 Vision-Language Models Fail to Generalize Across Modalities

A functional test for whether vision and language representations are alignable is to see if they can generalize out-of-distribution. Out-of-distribution generalization is important as an adapter between representations can act as a lookup table without relying on any shared underlying structure. For example, Mikolov et al. (2013b) found that, due to similar languages having some shared structure, a linear map between English and Spanish can translate words not in the map's training data. If English and Spanish had no shared structure then the linear map could only map from a word in English to its Spanish analogue, like a dictionary, without generalizing further. Thus, in some cases, representational misalignment directly causes a lack of generalization.

The multimodal analogue of machine translation would be to train a VLM as usual while omitting some concepts from its paired finetuning data, see Figure 3 (right). These concepts must be ones that the vision and language backbones recognize independently, as otherwise it would be surprising for the final VLM to suddenly recognize them. We discuss what it means for vision and language models to recognize or "know" of a concept in Appendix G. After training, the VLM's generalization can be tested by querying whether it recognizes the held-out concepts.[2]

Setup. We train a series of small generative and retrieval VLMs and large generative VLMs while omitting a few concepts from their datasets. Specifically, for the small-scale settings, we train

[2]This test is inspired by a thought experiment known as Molyneux's problem – briefly, could a blind person, upon regaining their sight, be able to visually distinguish what they previously only felt? If tested a few days after gaining their sight, the answer is, surprisingly, yes – see Held et al. (2011).

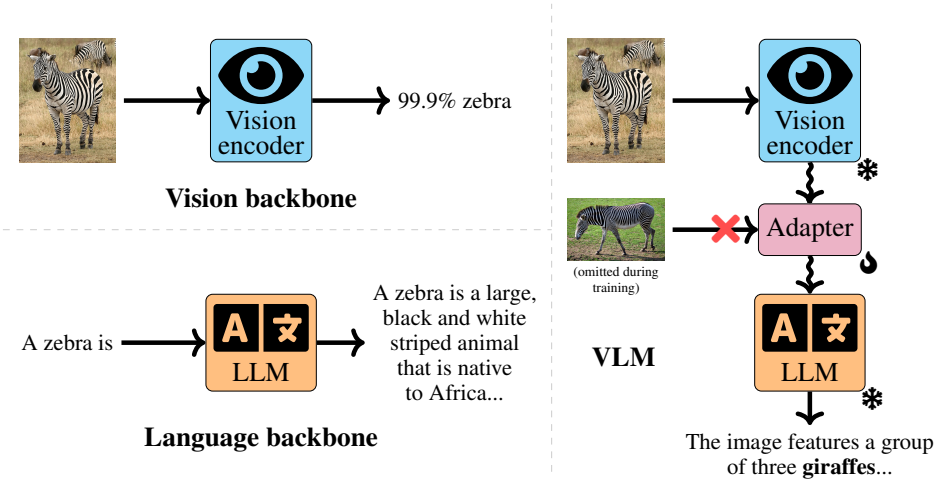


Figure 3: Pretrained vision and language models can each know what a zebra is in isolation, while a unified vision-language model fails if its adapter was not trained on zebra image-text pairs. Wavy lines denote feature vectors. Zebra images from Commons (2024).

adapters between frozen ViT-B/16 (Dosovitskiy et al., 2020) and GPT2-small (Radford et al., 2019) models for image captioning or retrieval, with Appendix F containing ablations and technical details.

For the large-scale setup, we train a LLaVA-1.5-7B model (Liu et al., 2023), with two important differences relative to how LLaVA is typically trained. First, for its second stage of training, we use low-rank adapters (LoRA) (Hu et al., 2022) instead of tuning the entire LLM to reduce the possibility of catastrophic forgetting. Second, LLaVA typically uses a CLIP vision backbone, which already received language supervision during its pretraining and is already multimodal. Instead, we use a ViT-L/16@384 model that is trained for ImageNet classification. This ViT only saw images during its pretraining and has the same number of parameters and tokens as the default vision backbone.

Concept filtering and detection. In all settings, the held-out concepts are filtered from the finetuning datasets, with details in Appendix H. Specific concepts were selected due to being simple, synonymless, and are verifiably known by the models (e.g., all are ImageNet classes, so an image classifier clearly recognizes them). To measure how well a model detects a held-out concept after training, we construct a new test set consisting of concept images. Images containing the concepts are selected by querying ImageNet validation data using a CLIP model and filtering the resulting examples manually, yielding $\sim$100 images per concept. For generation, a concept is detected if it appears in the generated answer, whereas for retrieval we benchmark the model on text-based image retrieval over 1000 images, one from each ImageNet class, including those of the held-out concepts.

Table 2: Held-out concept detection accuracy across settings. In all cases, the trained VLMs fail to detect the held-out concepts when they are not included in their paired image-text finetuning data. Large-scale generation refers to LLaVA-1.5-7B. Values in brackets denote bootstrapped 95% confidence intervals across per-concept accuracies. †(NLP Connect, 2022)

Setup (# Parameters)	Held-out Concept Detection Accuracy
Retrieval (210M)	
Without concepts	3.8% [0.2%, 9.2%]
Full dataset	84.3% [58.9%, 98.4%]
CLIP (ViT-B/16)	79.2% [46.5%, 97.6%]
Generation (210M)	
Without concepts	1.0% [0.0%, 3.0%]
Full dataset	84.0% [57.9%, 98.6%]
End-to-end finetuning†	87.8% [66.3%, 99.8%]
Large-scale generation (7B)	
Without concepts	21.5% [11.9%, 32.8%]
Full dataset	87.3% [80.0%, 94.1%]
CLIP backbone	95.3% [89.8%, 99.3%]

Results. Table 2 shows that VLMs trained without the held-out concepts struggle to detect them compared to models trained on the regular full datasets. To ensure the models are properly trained, each section's last row compares to a model trained elsewhere, showing they achieve similar performance.



(a) The image features **a large, round pie** with a crust, filled with a mixture of...



(b) The image features a woman holding **a colorful kite** in a field.



(c) The image features a bathroom sink with **a toothbrush** and a cup of water placed on it.

Figure 4: Image descriptions given by a LLaVA-1.5-7B model trained without the held-out concepts on images containing them when prompted "What is in this image?". The corresponding concepts are (a) pizza, (b) umbrella, and (c) toothbrush. The model fails to detect them for (a) and (b), while it succeeds for (c). See Appendix I for more examples.

Figure 4 contains qualitative examples of descriptions generated by the LLaVA model trained without the held-out concepts.

For large-scale generation—although the LLaVA model trained without the concepts attains much worse performance than the one trained with them—it is still higher than the small-scale experiments' accuracies. We speculate why this may be in Appendix J.

5 Representationally Aligned VLMs Exhibit Fewer Failures

The generalization failure described in the last section stems from representational misalignment. Representational misalignment is interesting in that, without explicit efforts to induce better alignment, misalignment should persist. As many VLM failures persist regardless of model scale, this raises the question whether additional failures may stem from representational misalignment as well (Diwan et al., 2022; Rahmanzadehgervi et al., 2024; Vo et al., 2025).

To test if representational alignment leads to fewer failures, we train 37 LLaVA models with different vision backbones and benchmark them on prediction tasks where VLM failures are common. All vision backbones use the same vision transformer architecture but have different pretraining objectives, such as image classification, multimodal contrastive losses, and so on. Thus, the only difference between models is the backbone's resulting representations. For additional details see Appendix K.

To test whether improved representational alignment results in fewer failures we use the following datasets: **Winoground** measures matching accuracies between images with syntactically similar but semantically different captions (Thrush et al., 2022). **VLMBias** tests whether VLMs can count objects in counterfactual images (e.g., the number of legs on a three-legged chicken) (Vo et al., 2025). **VLMBlind** is a suite of simple geometric tasks like tracing paths in a simplified subway map (Rahmanzadehgervi et al., 2024). **MMVP** tests if a model can differentiate between pairs of images which are "CLIP-blind", images which are similar per CLIP but dissimilar per other models (Tong et al., 2024).

Additionally, as a control, we test whether better representational alignment is correlated with better performance on datasets requiring the model to have properties orthogonal to representation alignment, such as reasoning. **Bongard** is a visual dataset consisting of geometric visual analogies that require reasoning about the relation between different sets of shapes (Wüst et al., 2024). **OmniSpatial** is a visual question answering spatial reasoning dataset (Jia et al., 2026).

Figure 5 shows that models with relatively higher representational alignment between their vision and language backbones tend to perform better on benchmarks that showcase VLM failures. Notably, Figure 5 (right column) shows that the correlation is not simply due to some vision backbones having better representations, as for the reasoning datasets the correlation is close to zero.

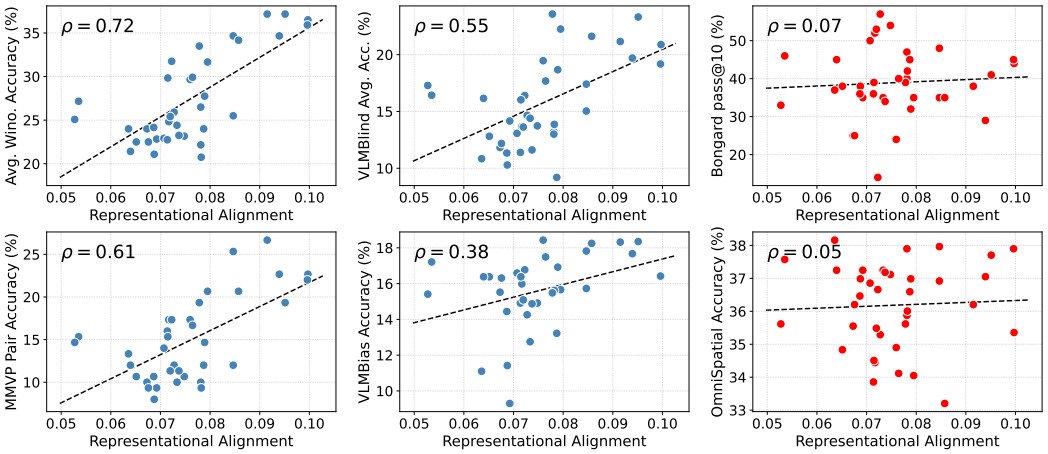


Figure 5: **Vision/language backbone representational alignment is correlated with better performance on benchmarks showcasing VLM failures** (left four plots). Each dot represents a LLaVA model trained with a different vision backbone. Representational alignment is measured using the mutual nearest neighbours similarity metric of Huh et al. (2024). (right two plots) show that on datasets which are difficult due to the model not having other capabilities, e.g. reasoning, the correlation is much smaller.

6 Discussion

In this paper we provide evidence that existing vision and language representations are linearly unalignable. Although these results are only over existing models, they are expected to hold regardless of model size due to the underlying reasons *why* the representations are misaligned. In Section 3, we find that the misalignment stems from the vision and language modalities having inherently different information; so what is similar in one may be dissimilar in another, and analogies may not carry over between modalities. As these differences are due to the data—not the models or training paradigm—they are expected to persist across all scales.

This implies that representations in different modalities should not converge—disproving a strong version of the Platonic Representation Hypothesis. Weaker versions of the hypothesis, where partial alignment is posited, can (and likely do) hold, consistent with prior work. Indeed, the weak trend in Figure 2 (left), where the vision models' retrieval accuracy increases with scale, supports this conclusion.

A natural question is whether alignment is possible using more complex, nonlinear adapters between modalities. As the misalignment is due to the data, better adapters could not lead to alignment, as it is impossible for the adapter to regain lost information. For example, captioning a picture filled with a single color as "green" does not convey the color's shade or hue.

More practically, representational misalignment has functional implications. We hypothesize that many noted VLM failure modes are due to representational misalignment, explaining why they persist in spite of continued improvements in model performance. The generalization failure demonstrated in Section 4 and the correlation between having fewer failures and representational alignment in Section 5 both support this hypothesis.

There are several caveats to our conclusions, with a broader discussion deferred to Appendix L. Most notably, our results only intervene in one direction – showing that representational misalignment can lead to failures – and do not yet show that fixing the misalignment directly reduces the failures. While there are some works which broadly show that better alignment leads to improved performance (Ma et al., 2024; Yamaguchi et al., 2025; Gan et al., 2026), finding a proper fix is an important avenue for future work. Moreover, although we show that representational alignment is linked to better results on datasets exhibiting VLM failures, this only be a component and not the failures' sole cause. Some failures likely stem from other causes, such as a lack of reasoning capabilities.

AI USE STATEMENT

Language models, typically a combination of Claude and ChatGPT models, were used for writing some of the experiment's code, checking all proofs, helping find the bounds connecting representational similarity to linear alignability in Appendix B when prompted to do so, writing the first version of the caption for Figure 8, generating code for plots, latex formatting, generating bibtex for the finetuned ViTs in Section 5, ensuring this statement is sufficiently thorough, helping make Appendix E.1, and helping think of different phrasings for specific sentences. No datasets were generated using any form of generative AI, nor were any language models used to directly write any of the paper's text. However, as VisLingA is downloaded from Wikimedia images, it may contain AI generated images created by others. All proofs and code were manually checked by the authors and verified to pass some basic sanities. Any remaining errors or imprecisions in the final manuscript are the full responsibility of the authors.

ETHICS STATEMENT

We do not foresee any significant ethical consequences of this work. As far as we are aware, VisLingA does not contain concepts that would likely pose any safety risks. However, as it contains images from Wikimedia, it contains trace amounts of sensitive publicly available images, such as partial nudity and historical violence. Creator metadata for all images in VisLingA is kept, with only images with CC0 or CC-BY copyright licenses being used.

REPRODUCIBILITY STATEMENT

Details necessary to reproduce all experiments are given in the main text and throughout the appendices. All code and datasets will be publicly released.

REFERENCES

Abhilash. ViT Face Emotion Detection. `https://huggingface.co/abhilash88/face-emotion-detection`, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences. *arXiv preprint arXiv:2506.02095*, 2025.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Wikimedia Commons. Zebra — wikimedia commons, the free media repository, 2024. URL https://commons.wikimedia.org/wiki/Zebra.

Adrián Csiszárik, Péter Kőrösi-Szabó, Akos Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. *Advances in Neural Information Processing Systems*, 34:5656–5668, 2021.

Dark Coder. vit-chest-xray. https://huggingface.co/codewithdark/vit-chest-xray, 2025.

Grégoire Dhimoïla, Victor Boutin, Agustin Martin Picard, Thomas Fel, and Thomas Serre. A unifying framework for concept-based representational similarity. *arXiv preprint arXiv:2606.09653*, 2026.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pp. 3519–3530, 2016.

Chester Enright. rorshark-vit-base. https://huggingface.co/amunchet/rorshark-vit-base, 2023.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Yarin Gal. The robustness of massive models pre-trained on web-scraped data is tightly tied to the way their data was curated. Technical report, Oxford Applied and Theoretical Machine Learning (OATML) Group, University of Oxford, June 2021. URL https://oatml.cs.ox.ac.uk/blog/2021/06/27/web-scraped-harmful.html. Blog post titled "Why I Wouldn't Trust OpenAI's CLIP to Drive My Car".

Yulu Gan, Kaiya Ivy Zhao, Tomaso Poggio, and Phillip Isola. Seeing helps reasoning in language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7080–7090, 2026.

Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. Vla-0: Building state-of-the-art vlas with zero modification. *arXiv preprint arXiv:2510.13054*, 2025.

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uz0gAAYydl.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. *arXiv preprint arXiv:2510.08492*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Zoe Wanying He, Sean Trott, and Meenakshi Khosla. Seeing through words, speaking through pixels: Deep representational alignment between vision and language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35645–35660, 2025.

Richard Held, Yuri Ostrovsky, Beatrice de Gelder, Tapan Gandhi, Suma Ganesh, Umang Mathur, and Pawan Sinha. The newly sighted fail to match seen with felt. *Nature neuroscience*, 14(5): 551–553, 2011.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Dmytro Iakubovskyi. ai_vs_real_image_detection. `https://huggingface.co/dima806/ai_vs_real_image_detection`, 2023a.

Dmytro Iakubovskyi. closed_eyes_image_detection. `https://huggingface.co/dima806/closed_eyes_image_detection`, 2023b.

Dmytro Iakubovskyi. deepfake_vs_real_image_detection. `https://huggingface.co/dima806/deepfake_vs_real_image_detection`, 2023c.

Dmytro Iakubovskyi. facial_emotions_image_detection. `https://huggingface.co/dima806/facial_emotions_image_detection`, 2023d.

Dmytro Iakubovskyi. man_woman_face_image_detection. `https://huggingface.co/dima806/man_woman_face_image_detection`, 2023e.

Dmytro Iakubovskyi. fairface_age_image_detection. `https://huggingface.co/dima806/fairface_age_image_detection`, 2024a.

Dmytro Iakubovskyi. fairface_gender_image_detection. `https://huggingface.co/dima806/fairface_gender_image_detection`, 2024b.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, volume 2026, pp. 35634–35670, 2026.

Om Khangaonkar, Hadi J Rad, and Hamed Pirsiavash. Multimodal language models cannot spot spatial inconsistencies. *arXiv preprint arXiv:2604.00799*, 2026.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Wenxuan Ma, Shuang Li, Lincan Cai, and Jingxuan Kang. Learning modality knowledge alignment for cross-modality transfer. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=lmiurzioja.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple word2vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5030–5047, 2024.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Harry Ng. vit-base-patch16-224-blur_vs_clean. https://huggingface.co/harrytechiz/vit-base-patch16-224-blur_vs_clean, 2023.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855, 2022.

NLP Connect. vit-gpt2-image-captioning (revision 0e334c7), 2022. URL https://huggingface.co/nlpconnect/vit-gpt2-image-captioning.

Amanda Nurul. watermark_detector. https://huggingface.co/amrul-hzz/watermark_detector, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, pp. 39643–39666. PMLR, 2024.

Szymon Pawlonka, Mikołaj Małkiński, and Jacek Mańdziuk. Bongard-rwr+: Real-world representations of fine-grained concepts in bongard problems. *arXiv preprint arXiv:2508.12026*, 2025.

Steven T Piantadosi, Dyana CY Muller, Joshua S Rule, Karthikeya Kaushik, Mark Gorenstein, Elena R Leib, and Emily Sanford. Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9):844–856, 2024.

prithivMLmods. Deep-Fake-Detector-v2-Model. https://huggingface.co/prithivMLmods/Deep-Fake-Detector-v2-Model, 2025. Initial release 21 Mar 2024; updated 02 Feb 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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*, pp. 8748–8763. PMLR, 2021.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

Rahul Ramachandran, Ali Garjani, Roman Bachmann, Andrei Atanov, Oğuzhan Fatih Kar, and Amir Zamir. How well does gpt-4o understand vision? evaluating multimodal foundation models on standard computer vision tasks. *arXiv preprint arXiv:2507.01955*, 2025.

James O Ramsay, Jos ten Berge, and George PH Styan. Matrix correlation. *Psychometrika*, 49(3): 403–423, 1984.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Anna Rogers, Aleksandr Drozd, and Bofang Li. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017)*, pp. 135–148, 2017.

Prithiv Sakthi. Age-Classification-SigLIP2. https://huggingface.co/prithivMLmods/Age-Classification-SigLIP2, 2025a.

Prithiv Sakthi. Gender-Classifier-Mini. https://huggingface.co/prithivMLmods/Gender-Classifier-Mini, 2025b.

Prithiv Sakthi. Watermark-Detection-SigLIP2. https://huggingface.co/prithivMLmods/Watermark-Detection-SigLIP2, 2025c.

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 27836–27864, 2025.

Jai Sharma. UL_base_classification. https://huggingface.co/sharmajai901/UL_base_classification, 2024a.

Jai Sharma. UL_bedroom_classification. https://huggingface.co/sharmajai901/UL_bedroom_classification, 2024b.

Jai Sharma. UL_interior_classification. https://huggingface.co/sharmajai901/UL_interior_classification, 2024c.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Damian Smith, Harvey Mannering, and Antonia Marcu. Functional alignment can mislead: Examining model stitching. In *Forty-second International Conference on Machine Learning*, 2025.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Christopher J Cueva, Erin Grant, Iris Groen, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Anke Tang, Li Shen, Yong Luo, Enneng Yang, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. Fusionbench: A comprehensive benchmark of deep model fusion. *Journal of Machine Learning Research*, 2025. arXiv:2406.03280.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. *arXiv preprint arXiv:2502.16282*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15395–15404, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Vishaal Udandarao. Understanding and fixing the modality gap in vision-language models. *Master's thesis, University of Cambridge*, 32, 2022.

Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M Asano. Same content, different answers: Cross-modal inconsistency in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8781–8790, 2026.

Rahul Vishwakarma. Human-Action-Recognition-VIT-Base-patch16-224. `https://huggingface.co/rvv-karma/Human-Action-Recognition-VIT-Base-patch16-224`, 2023.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Sophie L Wang, Phillip Isola, and Brian Cheung. Words that make language models perceive. *arXiv preprint arXiv:2510.02425*, 2025b.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *Advances in Neural Information Processing Systems*, 36:35331–35349, 2023.

Antonia Wüst, Tim Woydt, Lukas Helff, Inga Ibs, Wolfgang Stammer, Devendra S Dhami, Constantin A Rothkopf, and Kristian Kersting. Bongard in wonderland: Visual puzzles that still make ai go mad? *arXiv preprint arXiv:2410.19546*, 2024.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4256–4266, 2025.

Eunice Yiu, Maan Qraitem, Anisa Noor Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. Kiva: Kid-inspired visual analogies for testing large multimodal models. *arXiv preprint arXiv:2407.17773*, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Tyler Zhu, Tengda Han, Leonidas Guibas, Viorica Pătrăucean, and Maks Ovsjanikov. Dynamic reflections: Probing video representations with text alignment. *arXiv preprint arXiv:2511.02767*, 2025.

A Related Work

Adding modalities can harm performance. The generalization failure and results on representational misalignment here demonstrate that when representations are unalignable then stitching them together leads to worse performance, expressed as a lack of generalization. While this work focuses on tasks with language outputs, Ramachandran et al. (2025) similarly find that state-of-the-art VLMs perform worse than vision-only models on standard vision tasks, although without interpreting why. Our work fills this gap.

A converse conclusion would be that removing modalities could improve generalization. This was observed, albeit not understood, in Goyal et al. (2025). Goyal et al. (2025) use a VLM for a robot's vision-language-action model (VLA) without explicitly mapping into action-space, instead conveying actions as text. Their approach outperforms prior state-of-the-art methods that explicitly add the action modality.

There are cases when adding modalities would be expected to improve performance, but those likely require only in-distribution generalization. When the additional modalities enable training from large-scale data (Radford et al., 2021) or better supervision, e.g. using multi-task learning (Wang et al., 2025a), it is likely that in-distribution performance will improve.

Representational misalignment. Many works observe some form of multimodal representational misalignment, but often attribute it to a lack of scale (as would be expected per the Platonic Representation Hypothesis (Huh et al., 2024)) or do not interpret it; see Shu et al. (2025) for a survey.

Notably, Liang et al. (2022) note there being a "modality gap" between vision and language representations in CLIP-like contrastive learning, attributing it not to the inherent information difference in the modalities but to model initialization and optimization. This viewpoint is contested (Udandarao, 2022), with Schrodi et al. (2025) arguing, similarly to this work, that the modality gap is due to inherent information differences in the data.

Several prior works note caveats to the Platonic Representation Hypothesis. Gröger et al. (2026) note that representations can seem to converge simply due to confounders, e.g. wider models can have higher representational similarity. The experiments in Section 5 are designed so all models have the same architecture, circumventing this pitfall. Li et al. (2024) speculate that some representations might be inherently unalignable, such as vision and language, with our work providing evidence this is indeed the case.

Analogies in representations. Analogies have been seen to exist in representations since the seminal work of Mikolov et al. (2013a), with many contemporary works assume or rely on analogies in representations (Trager et al., 2023; Turner et al., 2023; Wang et al., 2023; Piantadosi et al., 2024; Merullo et al., 2024). Analogies have been argued to arise due to concepts being represented as linear directions in representation spaces ("the Linear Representation Hypothesis", (Park et al., 2024)). However, this viewpoint has some known limitations, especially when analogies are used for retrieval (Drozd et al., 2016; Rogers et al., 2017).

VLM failures. VLMs exhibit many simple failures, some of which are not analyzed here. Yuksekgonul et al. (2022) demonstrate instances where VLMs can act as bags-of-words and propose a contrastive learning method to mitigate this behaviour. Tong et al. (2024) demonstrate VLMs giving systematically wrong answers over certain curated visual question answering datasets, likely due to their CLIP vision backbones. They explore the design space of VLMs and propose changes that improve performance. Pawlonka et al. (2025) show that VLMs struggle perceiving fine-grained concepts in everyday real-world images. Khangaonkar et al. (2026) find that VLMs struggle detecting spatial inconsistencies between sets of images, e.g. a rotated chair. van Sprang et al. (2026) show that some reasoning inconsistencies likely arise due to representational misalignment, which could imply that the analysis in section 5 applies also to some reasoning tasks. Campbell et al. (2024) posit that various VLM failures are due to the binding problem, where models fail to associate together different features of an object.

B Bounds On Some Representational Alignment Metrics

In this appendix we show how if two sets of representations are linearly alignable only to some given extent then some representational alignment metrics are bounded in some sense by the degree of alignment. It is easiest showing these results when using different measures of linear alignability for different similarity metrics, due to how different similarity metrics measure different properties of the representations. However, any alignability metrics we use being zero implies the two representations are linearly alignable.

Notation. We first introduce some notation, mostly following that of Kornblith et al. (2019). Intuitively, given corresponding sets of representations $Y \in \mathbb{R}^{n \times p_1}, X \in \mathbb{R}^{n \times p_2}$,[3] X is linearly alignable to Y if there exists a weight matrix $W \in \mathbb{R}^{p_2 \times p_1}$ such that $XW = Y$. X is a matrix consisting of n representations with p_2 features, with Y being another such matrix but with p_1 features. Following Kornblith et al. (2019), we assume that $p_1 \leq p_2$ and that X, Y have been centred, or in other words, have zero mean over their columns.

CCA. Canonical correlation analysis (CCA) finds linear transformations that maximize the correlations between two sets of representations (Hotelling, 1992; Morcos et al., 2018). One similarity index given by CCA is Yanai's GCD measure (Ramsay et al., 1984; Kornblith et al., 2019), which between X and Y is $R^2_{\text{CCA}} \coloneqq \frac{||Q_Y^T Q_X||_F^2}{p_1}$, where Q_Y, Q_X are matrices with orthonormal columns spanning the columns of Y and X respectively.

CKA. Centred kernal alignment (CKA) is an extension of some prior representational similarity indices, which is notably invariant to orthogonal transformations but not to linear maps in general

[3]Note we assume X has at least as many features as Y while Kornblith et al. (2019) assume the opposite; this is purely notational but makes some proofs more intuitive later.

(Kornblith et al., 2019). Kornblith et al. (2019) showed that in many cases CKA better identifies intuitively corresponding representations than other similarity indices. Formally, CKA is defined as $\text{CKA} := \frac{||Y^T X||_2^F}{||X^T X||^F ||Y^T Y||^F}$.

Limited linear alignability bounds CCA. Arguably the simplest alignability metric is $l := \min_W ||Y - XW||_F^2$, where $l = 0$ if there exists a weight matrix such that $Y = XW$ and is nonzero otherwise. Here we show how l is bounded if X, Y are not linearly alignable, and accordingly leads to a bounded CCA. Note that we do not assume X, Y, or W have any special form, beyond that X has at least as many features as Y.

Note that Y decomposes into a matrix contained in X's column space and a matrix orthogonal to it. $l = 0$ if Y is wholly contained in X's column space, as then it is a linear combination of X's columns, in which case X and Y are linearly alignable. More generally, denote by P_X a projection onto X's column space, such that $Y = P_X Y + (I - P_X)Y$. Thus, note that $l = \min_W ||Y - XW||_F^2 = \min_W ||(I - P_X)Y + P_X Y - XW||_F^2$. As $P_X Y - XW$ is contained in X's columns while $(I - P_X)Y$ is orthogonal to it, from Pythagoras' theorem we have that $l = ||(I - P_X)Y||_F^2 + \min_W ||P_X Y - XW||_F^2 = ||(I - P_X)Y||_F^2$, where the second term is zero as $P_X Y$ is in the column space of X, which is spanned by XW.

Thus, $l = ||(I - P_X)Y||_F^2$. To relate Y to how it appears in CCA's similarity index, note that each of its columns is a vector in its column space. Intuitively, if $q_1, ..., q_{p_1}$ is an orthonormal basis for Y's column space, then each vector in the column space can be decomposed as $y = \sum_{i=1}^{p_1} (q_i^T y) q_i$. In matrix form, this gives the identity $Y = Q_Y Q_Y^T Y$, with Q_Y being an orthonormal basis column for Y's column space. More generally, for any Z we have that $P_Z = Q_Z Q_Z^T$, where P_Z is the projection matrix onto Z's column space and Q_Z is a matrix of that space's orthonormal basis.

Now, note that as $||AB||_F^2 \leq ||A||_F^2 ||B||_F^2$, we have that $l = ||(I - P_X)Y||_F^2 = ||(I - P_X)Q_Y Q_Y^T Y||_F^2$, as Q_Y acts as the identity on Y. Using the inequality, we have that $l \leq ||(I - P_X)Q_Y||_F^2 ||Q_Y^T Y||_F^2 = ||(I - P_X)Q_Y||_F^2 ||Y||_F^2$, with the last step being due to $Q_Y Q_Y^T$ acting as the identity on Y, where $||Q_Y^T Y||_F^2 = \text{Tr}(Y^T Q_Y Q_Y^T Y) = \text{Tr}(Y^T Y) = ||Y||_F^2$.

We now connect this quantity to CCA. As the Frobenius norm is invariant to orthonormal matrices, note that $||Q_X^T Q_Y||_F^2 = ||Q_X Q_X^T Q_Y||_F^2 = ||P_X Q_Y||_F^2$. We now note that this is the complementary part of a projection term that bounds l, namely the $||(I - P_X)Q_Y||_F^2$ term. Thus, using Pythagoras' theorem, $||(I - P_X)Q_Y||_F^2 + ||P_X Q_Y||_F^2 = ||Q_Y||_F^2 \leq p_1$, with the last step being due to Q_Y being an orthonormal column matrix, with the inequality being tight when Y is full rank. Plugging these results together, we have that:

$$R_{\text{CCA}}^2 = \frac{||Q_X^T Q_Y||_F^2}{p_1} = \frac{||Q_Y||_F^2 - ||(I - P_X)Q_Y||_F^2}{p_1} \leq \frac{p_1 - ||(I - P_X)Q_Y||_F^2}{p_1} \leq 1 - \frac{l}{||Y||_F^2 p_1}.$$

Thus, if the linear alignability is bounded, where $l > \varepsilon > 0$, then $R_{\text{CCA}}^2 < 1 - \frac{\varepsilon}{||Y||_F^2 p_1}$ is bounded as well. Note that this bound is quite loose and for illustration only, as there are various places in the derivation where it could be made tighter.

Different similarity structures bound CKA. In section 3 we show that if two representations have the same similarity structure then there is a linear mapping between them. Here we expand on what happens when their similarity structures differ; namely, how larger differences between the similarity structures generally lead to lower CKA.

Denote by $K_X := XX^T, K_Y := YY^T$ the Gram matrices for X and Y respectively, which as per definition 3.4 denote each representation's similarity structure. Denote by $\hat{K}$ a normalized Gram matrix, where $\hat{K} := \frac{K}{||K||_F}$. Thus, a natural scale-invariant notion for the difference between two similarity structures is $d := ||\hat{K_Y} - \hat{K_X}||_F^2$. Linking this to CKA, note that $\text{Tr}(\hat{K}_Y^T \hat{K}_X) = \frac{\text{Tr}(K_Y^T K_X)}{||K_Y||_F ||K_X||_F} = \text{CKA}(X, Y)$. Thus, by expanding d we find that $d = ||\hat{K}_Y||_F^2 + ||\hat{K}_X||_F^2 - 2\,\text{Tr}(\hat{K}_Y^T \hat{K}_X) = 2(1 - \text{CKA}(X, Y))$.

Thus, if the difference between the normalized similarity structures is bounded, so $d > \varepsilon$, the CKA between them is bounded as well, where $\text{CKA}(X, Y) < 1 - \frac{\varepsilon}{2}$.

C PROOFS OF THEOREMS

We restate the theorems here for convenience. We believe essentially all of our theorems have standard linear algebra versions that are generally known but we did not find a source for some, so we provide short proofs for them instead. For ease of notation we drop the concepts in the notation, so $v_{X_i} = v_i$. We first prove Theorem 3.7 as Theorem 3.5 proof's relies on it, it being the weaker of the two.

Theorem 3.7 (Generalized analogies enable linear mappings.) *Let V and W have corresponding sets of concept representations, $\{v_i\}_{i=1}^N$ and $\{w_i\}_{i=1}^N$. A linear map T such that $\forall i : T(v_i) = w_i$ exists if and only if for all sets of scalars $a_i \in \mathbb{R}$ if $\sum_i a_i v_i = 0$ then $\sum_i a_i w_i = 0$.*

Proof. It is simple to see that having such a T is sufficient, as $\sum_{i=1}^N a_i v_i = 0 \Rightarrow T(\sum_{i=1}^N a_i v_i) = T(0) = 0$, and due to linearity $T(\sum_{i=1}^N a_i v_i) = \sum_{i=1}^N a_i T(v_i) = \sum_{i=1}^N a_i w_i$, which completes this part of the proof.

As for necessity, assume that $\forall a_i \in \mathbb{R} : \sum_{i=1}^N a_i v_i = 0 \Rightarrow \sum_{i=1}^N a_i w_i = 0$. We prove this using induction on N, the number of vectors. For the base case this is trivial, as only for $a_1 = 0$ the statement holds and the map $\forall a \in \mathbb{R} : T(av_1) = aw_1$ is clearly a valid linear transformation. For induction, assume the statement holds for $N-1$ and we wish to show it for N. If v_N is linearly dependent on $\{v_i\}_{i=1}^{N-1}$ then $\exists a_i \in \mathbb{R} : v_N = \sum_{i=1}^{N-1} a_i v_i$, so $-v_N + \sum_{i=1}^{N-1} a_i v_i = 0 \Rightarrow -w_N + \sum_{i=1}^{N-1} a_i w_i = 0 \Rightarrow w_N = \sum_{i=1}^{N-1} a_i w_i$. Thus, as $T(\sum_{i=1}^{N-1} a_i v_i) = T(v_N) = w_N$ the same linear transformation works here as well.

If v_N is linearly independent of $\{v_i\}_{i=1}^{N-1}$ then recall that for two vector spaces V, W there exists a linear transformation between a basis of V and any set of vectors in W. Assume without loss of generality that $\{v_i\}_{i=1}^{N-1}$ are linearly independent, as otherwise a vector could be removed from the set without altering its span and the induction hypothesis could readily be used with the reduced set and v_N. Thus, the set $\{v_i\}_{i=1}^N$ is linearly independent and the aforementioned lemma can be used for $V' \coloneqq \text{span}(\{v_i\}_{i=1}^N)$ and W. Therefore, such a T exists. If $V' \subset V$ then T can be extended to be over V by mapping basis vectors in V/V' to 0, thereby completing the proof. □

Theorem 3.5 (Same similarities enable linear mappings.) *Let V and W have corresponding sets of concept representations, $\{v_i\}_{i=1}^N$ and $\{w_i\}_{i=1}^N$. If $\forall i, j : v_i^T v_j = w_i^T w_j$ then there exists a linear map $T : V \to W$ such that $\forall i : T(v_i) = w_i$.*

Proof. We wish to show that $\forall a_i \in \mathbb{R} : \sum_i a_i v_i = 0 \Rightarrow \sum_i a_i w_i = 0$, thus allowing us to use Theorem 3.7. Let there be some set of a_i such that $\sum_i a_i v_i = 0$. Multiplying by $\sum_i a_i v_i$ and using that $v_i^T v_j = w_i^T w_j$, we have that $\| \sum_i a_i v_i \|^2 = \sum_{i,j} a_i a_j v_i^T v_j = \sum_{i,j} a_i a_j w_i^T w_j = 0$. Note that $\sum_{i,j} a_i a_j w_i^T w_j = \| \sum_i a_i w_i \|^2 = 0$, and as $\sum_i a_i w_i$ is of length zero it must be the zero vector. Thus, $\sum_i a_i w_i = 0$ and the proof is complete. □

It is easy to see that the resulting transformation $T(v_i) = w_i$ is an isometry over $\text{span}(\{v_i\})$ as $\forall a_i \in \mathbb{R} : \| \sum_{i=1}^N a_i v_i \|^2 = \sum_{ij=1}^N a_i a_j v_i^T v_j$ so due to the sets of concepts having the same similarity structure this equals $\sum_{ij=1}^N a_i a_j w_i^T w_j = \| \sum_{i=1}^N a_i w_i \|^2 = \| \sum_{i=1}^n a_i T(v_i) \|^2 = \| T(\sum_{i=1}^N a_i v_i) \|^2$, so distances are preserved. The corollary, where different similarity structures rule out there being an isometry between the representations, is straightforward to prove:

Corollary 3.6(Different similarity structures rule out isometric mappings) Assume V, W have corresponding sets of concept representations but different similarity structures. Thus, there cannot exist an isometric linear map $T : V \to W$ such that $\forall i : T(v_{X_i}) = w_{X_i}$.

Proof. Assume for the sake of contradiction that there exists such an isometric linear map. As T is an isometry and thus respects inner products, $\forall i, j : v_{X_i}^T v_{X_j} = T(v_{X_i})^T T(v_{X_j}) = w_{X_i}^T w_{X_j}$. However, as the two representations have a different similarity structure, $\exists i, j : v_{X_i}^T v_{X_j} \neq w_{X_i}^T w_{X_j}$, thereby yielding a contradiction and completing the proof. □

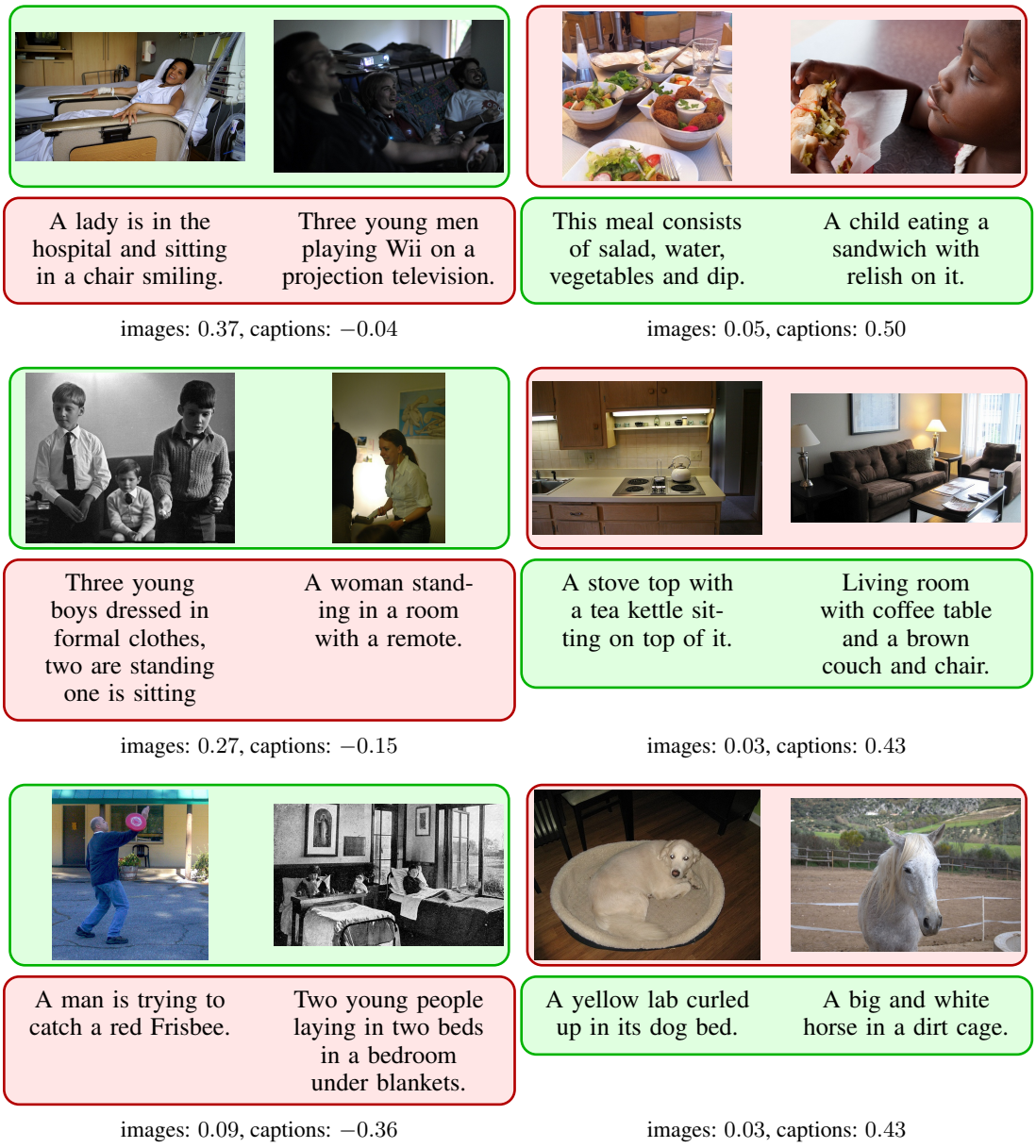


Figure 6: Additional examples of cases of similar images with dissimilar captions or vice versa. Beneath each pair is the cosine similarity between the images and the average embedding of all of their possible captions, with a random one of the captions being shown. Pairs are randomly uniformly from all pairs whose image and caption cosine similarities differ by more than 0.4.

D More Examples of Data With Different Similarity Structures

See Figure 6. Note that a small subset of examples seem to be due to errors in the semantic similarity models, e.g. the two images in the bottom left being classified as similar.

E Example Images from VisLinga

See Figure 7.

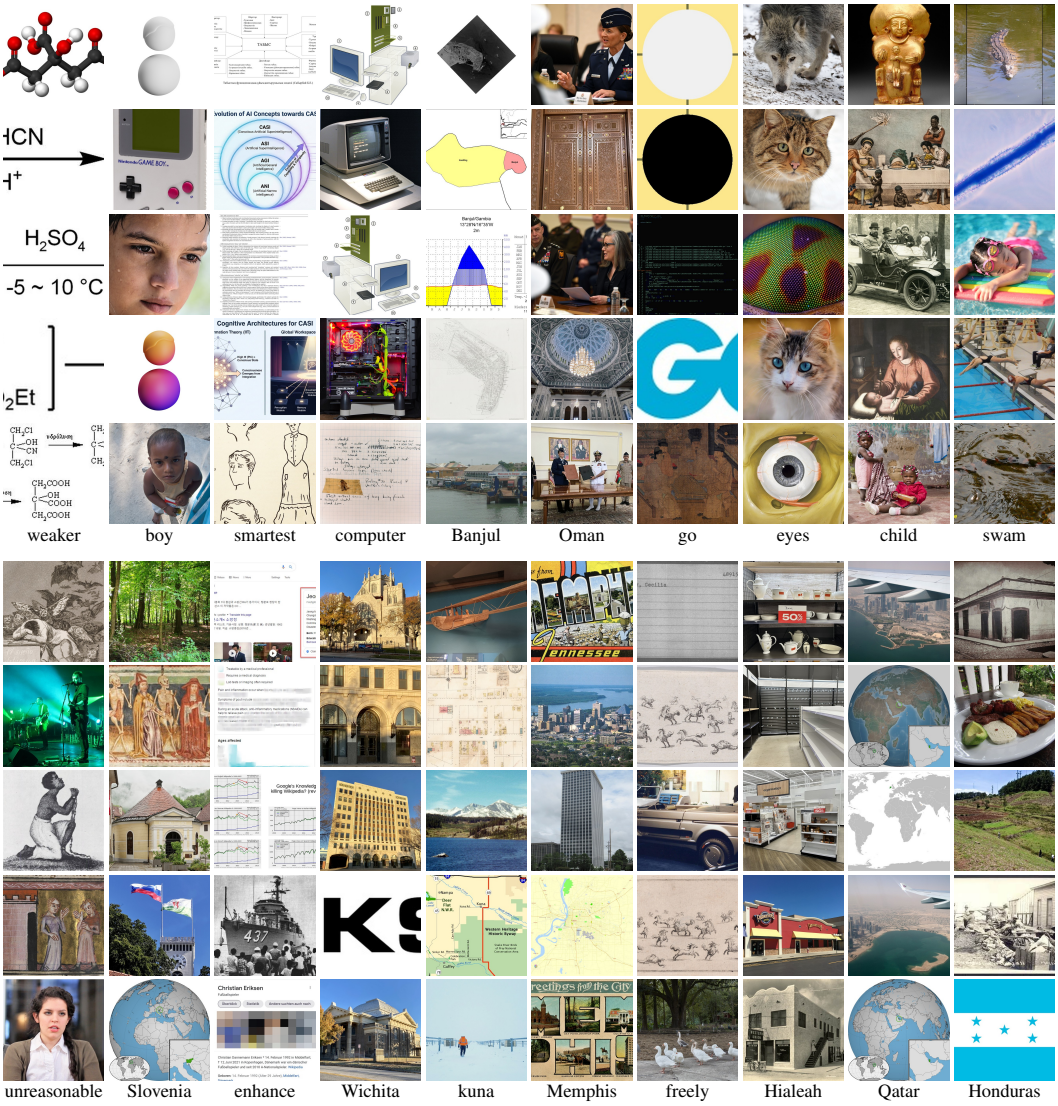


Figure 7: Randomly sampled VisLinga concepts and some of their images. Each column shows 5 randomly selected images for a randomly selected concept, named below the column. Images are center-cropped for display only and in practice have varying ratios.

E.1 Attribution for CC BY Images in Figure 7

All images in Figure 7 are from Wikimedia Commons and are in the public domain or have CC0/various CC BY licenses. The 46 images whose licenses require attribution are credited below, identified by the concept named under their column and their row (counted from the top of each block). Each row number links to the image's Commons page and each license name to the license. The other 54 images are in the public domain or released under CC0.

- *boy* (top): row 5: Vyacheslav Argenberg, CC BY 4.0.
- *smartest* (top): row 3: Serge Nicolas, Bernard Andrieu, Jean-Claude Croizet, Rasyid B. Sanitisio, Jeremy Trevelyan Burman, CC BY 3.0; row 5: J. E. Wallace Wallin, CC BY 4.0.
- *computer* (top): rows 1, 3: Gustavb, CC BY 2.5; row 2: Marcin Wichary, CC BY 2.0; row 4: Halicki, CC BY 3.0.
- *Banjul* (top): row 1: Capella Space, CC BY 4.0; row 3: Hedwig in Washington, CC BY 2.5; row 5: Niels Broekzitter, CC BY 2.0.
- *Oman* (top): rows 1, 2, 3, 4, 5: Chief, National Guard Bureau, CC BY 2.0.
- *eyes* (top): row 1: Christian Mehlführer, User:Chmehl, CC BY 2.5; row 2: Michael Gäbler, CC BY 3.0; row 3: Gribkov mikhail, CC BY 4.0; row 5: Patrick J. Lynch, medical illustrator, CC BY 2.5.
- *child* (top): row 5: Alfred Weidinger from Vienna, Austria; derivative work: Tomer T, CC BY 2.0.
- *swam* (top): row 1: MathKnight and Zachi Evenor, CC BY 4.0; row 3: Tommy Wong, CC BY 2.0; row 4: Dick Powell Photos Indy, CC BY 2.0; row 5: pintail, CC BY 4.0.
- *unreasonable* (bottom): row 2: kitmasterbloke, CC BY 2.0; row 5: Isaac Leicht, CC BY 3.0.
- *Slovenia* (bottom): row 1: Dibko, CC BY 4.0; row 3: Jakub Hałun, CC BY 4.0; row 4: User:Mateus2019, CC BY 3.0 de.
- *Wichita* (bottom): rows 1, 2, 3, 5: Myotus, CC BY 4.0.
- *kuna* (bottom): row 1: Brigade Piron, CC BY 3.0; row 3: David Ackerman, CC BY 2.0.
- *Memphis* (bottom): rows 1, 5: Steve Shook from Moscow, Idaho, USA, CC BY 2.0; row 3: Thomas R Machnitzki, CC BY 3.0.
- *Hialeah* (bottom): rows 1, 2, 3: Phillip Pessar, CC BY 2.0; row 4: Phillip Pessar from Miami, USA, CC BY 2.0.
- *Qatar* (bottom): rows 1, 4: Vyacheslav Argenberg, CC BY 4.0.
- *Honduras* (bottom): row 2: Luis Alfredo Romero, CC BY 4.0; row 3: Diario El Tiempo, CC BY 4.0.

F Small-Scale Experiments Details

For captioning, a linear layer is trained between the vision and language models as an adapter. For retrieval each has a linear layer trained on top of its last layer representations, by default using the class token for the vision head and averaging the post-adapter token representations for the language model. Different, more powerful adapters and representation aggregation schemes were not found to make a big difference. The training data is COCO-2014 (Lin et al., 2014).

To ensure the models are well trained optimal learning rates and weight decays were chosen by doing a grid search over learning rates and weight decays, spaced log-linearly in $10^{-5}, 10^{-4.5}, ...10^{-2}$ and $10^{-4}, 10^{-3}, ...10^{-1}$ respectively. Optimal retrieval hyperparameters were chosen based on the model's top-1 ImageNet classification accuracy over ImageNet's validation set. Generation hyperparameters were optimized on the BLEU score over COCO's validation set. Models were trained for 10 epochs with an early stopping patience of 5, with validation metrics computed once every 0.25 epochs.

Resulting optimal hyperparameters for all generative model configurations, including those referenced in the next section, are listed in Table 3. For retrieval a learning rate of 10^{-4} and weight decay of 10^{-2} was found to work well for all settings.

Adapter type	Learning rate	Weight decay
Linear	$10^{-2.5}$	10^{-2}
MLP	10^{-3}	10^{-2}
Attention	$10^{-3.5}$	10^{-3}

Table 3: Optimal learning rates and weight decays for different adapters for the small-scale generative experiments.

F.1 Different Adapters Yield Similar Results

To see if better adapters help we repeated the two small-scale experiments with two more powerful adapters – an MLP and an attention block between the tokens and learnable embeddings. The attention adapter has a single learnable token embedding for retrieval and 200 learnable embeddings for generation, where a ViT-B/16 has 197 vision tokens by default. For retrieval we also experiment with three different kinds of pooling – mean pooling, max pooling, and, specifically for the vision transformer, taking the class token. When the vision transformer uses the class token the language

model defaults to mean pooling. The attention adapter needs no pooling due to the cross attention already aggregating information.

In all cases, the linear adapter generalized best to the held-out concepts. We report not the held-out concept detection accuracy at the end of training but the best held-out concept detection accuracy throughout, which is a much more lenient metric. For generation the best concept detection accuracy throughout training is 7.8%, whereas for retrieval the pooling with a linear adapter did a bit better, getting a best concept detection accuracy throughout training of 11.1%. Test accuracies were measured after every 0.1 epochs.

G What Does it Mean for a Model to "Know" of a Concept?

For the vision models this is simple – all the concepts are ImageNet classes and the models are ImageNet classifiers that get high accuracies. For the language models, they can be prompted to complete texts such as "a `<concept>` is" or asked questions about it, thereby gauging some form of recognition.

To see whether the LLaVA backbone knows of the held-out concepts we ask it various questions about them, such as "What is a zebra?" and "What is a horse-like animal with black and white stripes?". Doing so with the LLaVA model's LLM backbone after it was trained showed that the language model by itself still recognizes the held-out concepts.

H Concept Filtering

The held-out concepts are "zebra", "banana", "pizza", "umbrella", and "toothbrush" for the small-scale experiments, with additionally "broccoli", "microwave", "strawberry", "castle", "envelope", "violin", "tractor", "flamingo", "printer", "volcano" for the large-scale experiments. The small-scale experiments use a smaller concept set as even the small scale models trained on the entire dataset fail to detect most of the additional concepts. To ensure the datasets are not overly reduced, for the large-scale experiments we train three models without the held-out concepts, each having only five concepts held-out.

To filter the held-out concepts from the finetuning data we remove all image-text pairs where a) the text contains the object's name or b) the image is likely to contain it. Image-based filtering is important to prevent misclassified data from hampering the experiment, e.g. zebra images accidentally captioned as giraffes. This is done using a CLIP ViT-L/14 model and filtering images with a similarity above a chosen threshold to the text "a photo of a `<concept>`", with the threshold found through manual tuning. For the small-scale experiments the threshold was 0.2 while for the large-scale ones it was 0.25. This reduces the datasets by at most 9%.

Specifically for the large-scale experiment we also filter the concept "tooth brush" as we find that it is occasionally not detected by the image-based filter. There may be other less direct forms of leakage we are unaware of that artificially boost the large-scale model's performance on detecting held-out concepts.

I More Qualitative Examples of LLaVA Generalization Failures

See Figure 8. Note that all descriptions are generated at $T = 0$, except for the successful banana example (o), as none of the $T = 0$ descriptions of the banana images mention a banana. Example (o) was obtained by sampling with $T = 0.2$ repeatedly.

J Discussion on Large-Scale Results

The large-scale model getting better held-out concept detection accuracies than the smaller models could mean that the failure described in Section 4 is partially alleviated by scale. However, there are other possible reasons a larger model could do better, such as LLaVA generally outputting longer completions or it potentially relying on its internal knowledge more so than what it sees, akin to the failure described by Vo et al. (2025). For example, the completion of "a bathroom with a toilet,



(a) The image features a plate of food, which includes **a sandwich and a salad**. The sandwich is placed on a pink plate...



(b) The image features a table with **a cake** on it, along with a birthday candle and a birthday card...



(c) The image features a large, delicious-looking **pizza** with various toppings, sitting on a plate.



(d) The image features a young boy standing in the rain, holding **a kite**.



(e) ...a man standing outside, holding a cell phone in his hand. He is also holding **a blue bag or purse**.



(f) The image features **a blue umbrella** sitting on the grass in a park.



(g) The image features a bathroom with a sink, mirror, and **a large collection of toiletries and personal care items**.



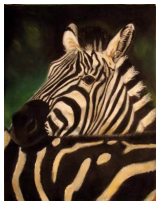
(h) The image features a woman **washing her hands** in a bathroom sink.



(i) ...bathroom with a mirrored cabinet, a sink, and a shelf. The shelf is filled with various items, including cups, **a toothbrush**, and a bottle.



(j) The image features **a giraffe** standing in a grassy field.



(k) The image features **a black and white striped giraffe**.



(l) The image features a close-up of **a zebra's face**, with its eyes and mouth visible.



(m) ...a man riding a bicycle with a cart attached to it, filled with **a large number of apples**...



(n) The image features a close-up of **two peeled oranges** sitting on a wooden table.



(o) The image features a large pile of yellow **bananas**, which are stacked on top of each other.

Figure 8: Additional examples of generalization failures from the LLaVA-1.5-7B model from Figure 4, trained without the held-out concepts, when prompted "What is in this image?". Rows correspond to the concepts pizza, umbrella, toothbrush, zebra, and banana respectively. For each concept we randomly sample two images whose description does not mention the concept (left, center) and one image where the concept was correctly detected (right). Note that this 2:1 ratio is purely illustrative and is not representative of the real detection rate per concept.

mirror, toothpaste, ..." would likely mention a toothbrush regardless of whether it actually exists in the image.

To test this, we benchmark the large-scale models when prompting them to only name the main object they see in the image, limiting them to output answers with at most 8 tokens. In this setting the models trained without the held-out concepts performed comparably well, getting an average detection accuracy of 19.6%. Thus, the large-scale model's success is not due to relying on the language model's knowledge while ignoring the vision model.

Another option is that there is some data leakage. While running these experiments, we initially filtered for "toothbrush" but not for "tooth brush", thereby having some examples slip by. As larger models are more sample efficient, these $\sim$20 extra image-text pairs resulted in toothbrushes being detected much more often. In the current setup, the large-scale models detect "pizza" much more than any other concept, although it is hard to say whether there is some data leakage we have not managed to find, the better detection performance is due to scale, or some mixture thereof. When training the model with the dataset visual similarity filtering threshold set to 0.2 instead of 0.25 models generally perform much worse, getting accuracies around half of those reported in Table 2, which lends credence to there possibly being some leakage.

K Details on Section 5 LLaVA Models

For the vision backbones we use 11 different standard pretrained ViT-B models and 26 finetunes thereof. ViT-B are used and not ViT-L, which is more typical for LLaVA, due to not all models having pretrained ViT-L variants. For the original set of 11 models, we use BEiT (Bao et al., 2021), CLIP models trained by OpenAI and on the Datacomp dataset (Radford et al., 2021; Gadre et al., 2023), DEiT (Touvron et al., 2021), DINO (Caron et al., 2021), DINOv2 (Oquab et al., 2023), MAE (He et al., 2022), SIGLip and SIGLip2 (Zhai et al., 2023; Tschannen et al., 2025), and ViTs trained on ImageNet 1k and 21k (Dosovitskiy et al., 2020). For the 26 finetunes we use 6 models from FusionBench (Tang et al., 2025) and various other models uploaded to HuggingFace, all with at least 400 downloads over the last month as of writing (Nurul, 2023; Iakubovskyi, 2023a;b;c;d;e; Enright, 2023; Ng, 2023; Vishwakarma, 2023; Sharma, 2024a;b;c; Iakubovskyi, 2024a;b; Dark Coder, 2025; Sakthi, 2025b; prithivMLmods, 2025; Sakthi, 2025a;c; Abhilash, 2025).

L Limitations

Representational similarity is generally a vague notion without a universally agreed upon definition, making it difficult to making decisive claims on representational similarity or alignability (Sucholutsky et al., 2023). While we attempt to make broad claims by showing connections between different measures of similarity (see section 2) and grounding results in theory where possible (section 3), these are naturally not exhaustive over the many notions of similarity. Additionally, results are believed to be general due to their underlying cause, namely inherent differences in the information content in different modalities. The difference in information content has been discussed in several prior works (Gal, 2021; Schrodi et al., 2025), with the results herein providing further evidence.

Most of the theoretical results are on linear alignability while in practice many VLMs are built using nonlinear adapters. Some results show that linear alignability bounds some notions of nonlinear similarity between representations (Appendix B), while others show that failures appear also given nonlinear alignment, namely the LLaVA experiments in sections 4 and 5, and the ablations in Appendix F.1.

Most results are circumstantial in supporting that many VLM failures likely arise due to representational misalignment. In this sense the results are consistent with the hypothesis but do not causally show that representational misalignment is what caused a failure. Results in section 4 are an important exception, as based on the theory they make a causal prediction which turns out to be correct – that if vision-language representations are misaligned then the generalization failure would occur.