VISION-LANGUAGE MODELS FAIL TO GENERALIZE ACROSS MODALITIES

Yonatan Gideoni[†] **Yoav Gelberg**[†] **Tim G. J. Rudner**[◦] **Yarin Gal**[†]
[†]OATML, University of Oxford [◦]University of Toronto
yg@robots.ox.ac.uk

ABSTRACT

Vision-language models exhibit many surprisingly simple failures, but why these failures occur remains unclear. We conjecture that their source is representational misalignment in the backbone's vision and language representations. We demonstrate a new generalization failure that would not occur if the representations were easily alignable, followed by a set of theory-grounded experiments further showing that the representations cannot be aligned using any linear transform. The representations are not expected to be better aligned with sufficient scale due to each modality containing inherently different information. Modern paradigms, such as reasoning or in-context learning, do not alleviate existing failures either. These results suggest that existing paradigms are incapable of preventing these failures and falsify a strong version of the Platonic Representation Hypothesis – that sufficiently powerful models trained in different modalities should converge to equivalent representations.

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 to identify 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 the same representations (Huh et al., 2024), with many 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). Research from over a decade ago showed that models within the same modality converge to similar representations, enabling inter-language retrieval (Mikolov et al., 2013b) and visual domain adaptation (Fernando et al., 2013), both of which translate between embeddings using only a linear map and show some out-of-distribution generalization. More recently, Moschella et al. (2023) showed that models in the same modality having similar representations enables zero-shot model stitching. Thus, if vision and language models converge to equivalent representations, would they generalize similarly to their unimodal counterparts?

Apparently not. We find that a VLM typically fails to recognize a concept its vision and language backbones know if it does not appear in its paired finetuning data (see Figure 1). To explain this generalization failure, we develop a simple framework for detecting when representations are equivalent up to linear transformations, with the representation spaces required to have either (1) the same sim-

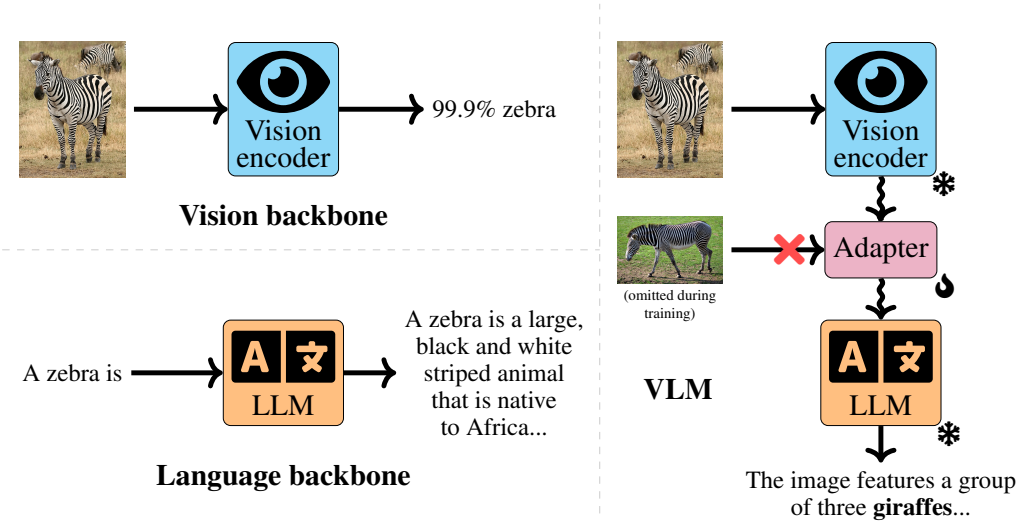


Figure 1: Pretrained vision and language models can each know what a zebra is in isolation, while a unified vision-language model fails if its adapter was not trained on zebra image-text pairs. Wavy lines denote feature vectors. Images from Commons (2024).

ilarity structure or (2) the same relationships between concepts. Vision and language representations are found to violate both conditions.

If VLM failures are caused by representational misalignment then orthogonal capabilities, like in-context learning (ICL) or reasoning, would not prevent the failures from occurring. Indeed, we find that more ICL examples and higher reasoning budgets do not alleviate the failures.

Altogether, these results support and lead us to broaden Diwan et al. (2022)'s conjecture, implying that **a diverse set of VLM failures likely stem from representational misalignment and will not be resolved solely by increasing model scale**. Our main contributions are:

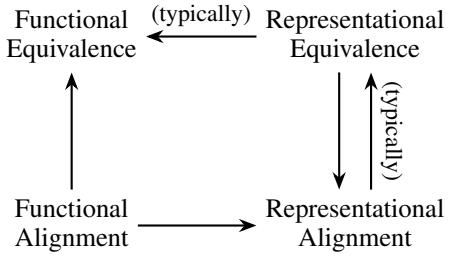
- Providing evidence a strong version of the Platonic Representation Hypothesis is false. This is evidenced by a new generalization failure (section 3) and experiments grounded in theory showing that vision and language representations are not equivalent and are likely to remain inequivalent regardless of scale (section 4).
- Showing that ICL and reasoning do not alleviate existing VLM failures (section 5), consistent with the hypothesis that many failures stem from representational misalignment.

2 Forms of Equivalence and Alignment

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 representations (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. The 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.



3 VLMs Fail to Generalize their Knowledge Across Modalities

A functional test to see 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, where a vision representation of a zebra is directly mapped to a zebra's language representation. 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.

A multimodal analogue would be to train a VLM as usual while omitting some concepts from its paired finetuning data, see Figure 1. These concepts must be ones 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 B. 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 scale generative and retrieval VLMs and large generative VLMs while omitting a few concepts from their datasets. Specifically, for the small-scale settings, we train 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 A containing ablations and technical details.

For the large-scale setup, we train a LLaVA-1.5-7B model (Liu et al., 2023). There are 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. For both settings the held-out concepts are filtered from the finetuning datasets, with details in Appendix C. Specific concepts were chosen due to appearing often in the datasets and they are simple, synonymless, verifiably known by the models (e.g., all are Im-

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

[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).

agenet classes, so an image classifier clearly recognizes them), and easily describable by analogy. For example, "a horse with black and white stripes" is a good approximation of a zebra. To measure how well a model detects a held-out concept after training, we construct a new test set consisting of images of the concepts. Images with a high correspondence to 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.

(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 2: 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).

Results. Table 1 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. Figure 2 contains qualitative examples of descriptions generated by the LLaVA model trained without the held-out concepts.

For the large-scale generation, although the LLaVA model trained without the concepts gets 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 D.

Table 1: 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. †(NLP Connect, 2022)

Setup (# Parameters)	Held-out Concept Detection Accuracy
Retrieval (210M)	
Without concepts	4.1%
Full dataset	86.6%
CLIP (ViT-B/16)	84.6%
Generation (210M)	
Without concepts	0.9%
Full dataset	88.3%
End-to-end finetuning†	91.3%
Large-scale generation (7B)	
Without concepts	23.6%
Full dataset	93.5%
CLIP backbone	98.0%

4 Why is intermodal alignment difficult?

To understand why aligning different modalities is hard it helps to formally understand why one would expect it to be easy and where this intuition falls short. The high-level intuition is that if both language and vision share the same analogical relations then it should be easy to map between their representations. 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 E. If v_X denotes some model's representation of concept X then an analogy is defined as follows.

Definition 4.1 (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.

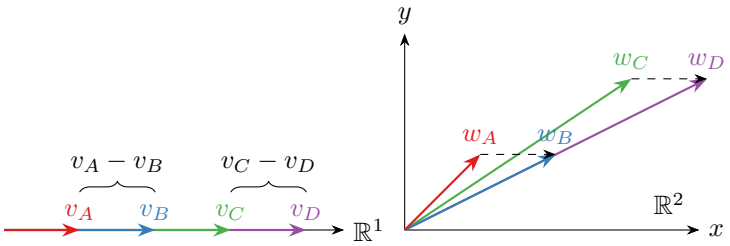
In practice, the analogy would only be approximate, so $v_A - v_B \approx v_C - v_D$. Throughout this section, take V, W to be two finite-dimensional inner product spaces, e.g. $\mathbb{R}^n$ and $\mathbb{R}^m$. One can intuitively think of them as the representations of two different models, potentially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

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 4.2 (Analogies are not sufficient 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 4.3 (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 following diagram illustrates this visually.



Example 4.3 can be equivalently interpreted as one representation having information which is missing in the other. Intuitively, the second index in the w vectors could be some additional 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 maps, is that the different representations have the same similarity structure. 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 texts "cat" and "dog".

Definition 4.4 (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 4.5 (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.

A weaker condition that still results in 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 4.6 (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 4.7 (Generalized analogies result in and require regular analogies)**.** *If the conditions of Theorem 4.6 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 4.6.*

In practice, text and vision have neither the same similarity structure nor the same analogies, respectively violating the conditions of both Theorem 4.5 and Theorem 4.6. Note that, as Theorem 4.6 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.

4.1 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 4.4), thus fulfilling the conditions of Theorem 4.5, two images with a given cosine similarity should have captions with approximately the same cosine similarity as well.

Figure 3: 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.

In practice, visual and lingual descriptions of the same scene can be similar in one modality but very different in another. Figure 3 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 Transformer's 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 3 occur with a non-negligible frequency. One way to measure their prevalence is by finding pairs where the cosine similarity between the images is much higher or lower than that between their captions, so the absolute difference in their cosine similarities is more than a relatively large threshold, e.g. 0.4. For about 2% of COCO image/caption pairs, the cosine similarity difference exceeds this threshold.

Table 2: Relative cosine similarity RMSEs for representations from different vision and language models. Similarities across different modalities are typically at least $3\times$ higher than those within modalities. All values are relative to Text,Text', where $\mathrm{CosSimRMSE}(\text{Text}, \text{Text'}) = 0.05$. 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

To quantitatively measure whether the two representations have the same similarity structure, we empirically compare their cosine similarities. Formally, for models f_1, f_2 outputting normalized vectors $f_1(x), f_2(x)$, we define $\mathrm{CosSimRMSE}(f_1, f_2) := \sqrt{\mathbb{E}[(f_1(x)^T f_1(y) - f_2(x)^T f_2(y))^2]}$, where the expectation is over the dataset. CosSimRMSE tests the extent to which two representations have the same similarity structure as per definition 4.4. 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 $\mathrm{CosSimRMSERatio}(f_1, f_2 | g_1, g_2) = \frac{\mathrm{CosSimRMSE}(f_1, f_2)}{\mathrm{CosSimRMSE}(g_1, g_2)}$, with g_1, g_2 representing two models in the same modality that form some lower bound. As Table 2 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.

Altogether, this section's results show that Theorem 4.5's condition does not hold, as vision and language have very different similarity structures. We now show that the conditions for Theorem 4.6 do not hold either, showing that typical vision and language representations are not representationally alignable.

4.2 Language and vision have different analogies

To see whether vision and language representations exhibit the same analogies we build on 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). 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 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 when performing retrieval 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.

Representations are computed using class tokens from DINOv3 base for the images and embeddings from Sentence Transformer's all-MiniLM-L12-v2 model for text.

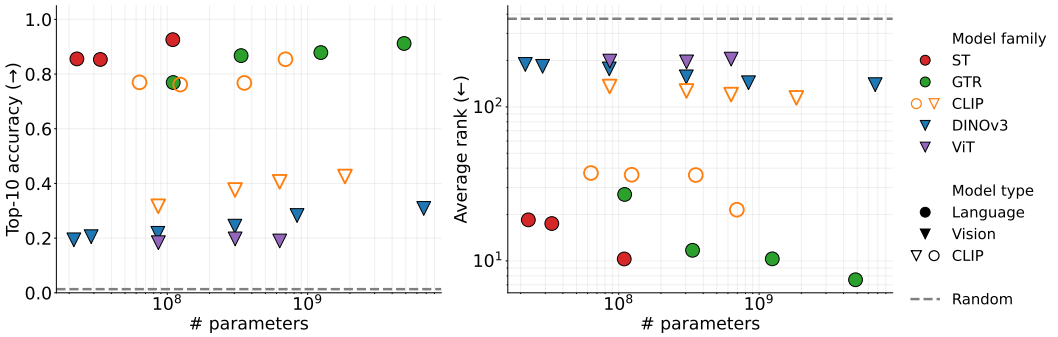


Figure 4: Vision models do not exhibit the same analogies that language models do. Left – top-10 analogy retrieval accuracies over VisLingA for several model families. The vision models (downwards triangles) do significantly worse than the text models (circles). Right – average retrieval rank. All vision models have a 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).

Results. Figure 4 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 4.6 cannot hold.

While it is not surprising that CLIP vision models perform similarly to other vision models, it is interesting that CLIP language models slightly underperform as well. The vision supervision they receive during training likely pulls their representations away from those of typical language-only models, thereby stopping them from recognizing some analogies.

5 ICL AND REASONING DO NOT MITIGATE VLM FAILURES

In-context learning (ICL) and reasoning have both been shown to significantly increase LLM and VLM performance over many tasks (Brown et al., 2020; Guo et al., 2025; Alayrac et al., 2022; Huang et al., 2025). For VLM failures which are due to vision-language representational misalignment, reasoning and ICL would not be expected to improve performance as they are orthogonal capabilities. If, however, for a given failure ICL or reasoning do improve performance then the failure likely stems not from representational misalignment.

We present preliminary evidence supporting this hypothesis by benchmarking a high-performing VLM, Qwen3 VL 8B Thinking (Bai et al., 2025), on Winoground given different reasoning budgets and numbers of ICL examples. Figure 5 shows that ICL and reasoning typically decrease performance. Results are for optimized prompts and ICL setup, with other tested variations resulting in worse and typically monotonically decreasing performance.

6 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

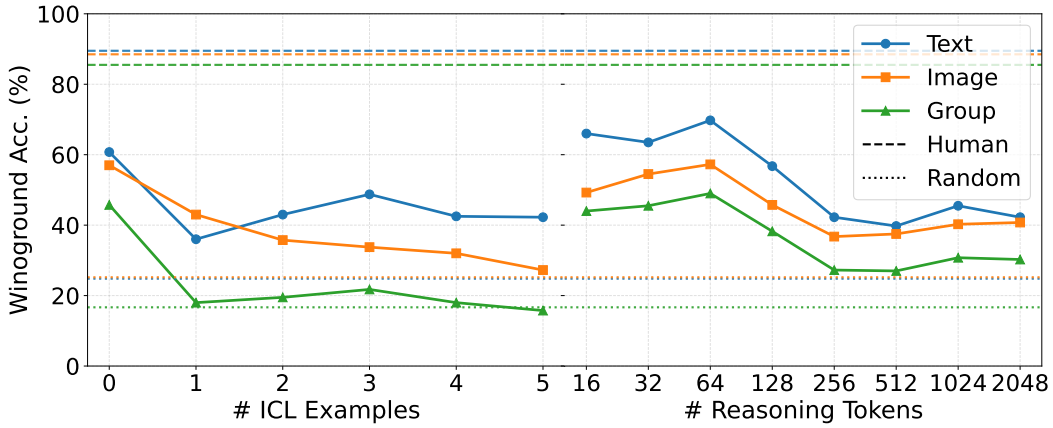


Figure 5: ICL and reasoning do not significantly affect performance on Winoground, a dataset where VLMs struggle reaching human performance.

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, although this viewpoint is contested (Udandarao, 2022).

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 as they have since been solved. 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.

7 Discussion

By demonstrating a functional generalization failure and a series of theory-grounded tests, this work has shown that existing vision and language representations are not representationally alignable and

therefore not equivalent. Although these results are only over existing models, they are expected to hold regardless of model size due to *why* the representations are misaligned. In section 4 we find that the misalignment stems from the modalities having inherently different information, so what is similar in one may be dissimilar in another, and analogies not carrying 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.

Thus, this implies that regardless of scale, representations in different modalities should not converge, disproving a strong version of the Platonic Representation Hypothesis. Weaker versions, where there is some partial alignment, can and likely do hold, consistent with prior work. Indeed, the weak trend in Figure 4 (left) supports this, where the vision models' retrieval accuracy increases with scale.

The 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, and why modern capabilities like in-context learning and reasoning do not mitigate the failures. The generalization failure presented in section 3 is another example of such a failure.

There are several caveats to these conclusions. On the link between representational misalignment and VLM failures, evidence here can only fail to refute the hypothesis, where the best evidence for it would be a solution to the representational misalignment. This will be pursued in future work. Regarding the evidence towards vision-language misalignment, although this work demonstrates the existence of an inherent vision-language gap, it does not discuss its extent. The degree to which vision and language representations may naturally converge is left unclear.

It would be interesting to devise methods which inherently have no representational misalignment, thus alleviating existing VLM failures while potentially broadly improving performance. One option for doing so is to treat both images and text as a single modality, processing them with masked next-token prediction. Doing so would make representational alignment unnecessary as there would be a single set of representations. This and other ideas will be pursued in future work.

ACKNOWLEDGEMENTS

We would like to thank Charlie Tan for suggesting the use of zebras as a typical example and to Dulhan Jayalath for reviewing a draft of this work.

REFERENCES

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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.

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.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pp. 2960–2967, 2013.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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.

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.

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.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

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

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

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.

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.

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.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 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`.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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 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.

A.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 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.

B 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.

C CONCEPT FILTERING

The held-out concepts are "zebra", "banana", "pizza", "umbrella", and "toothbrush". More concepts were not used so the datasets would be minimally altered.

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.

D 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 3 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, 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 did not filter for "tooth brush", only "toothbrush", thereby having some examples slip by. As larger models are more sample efficient, these $\tilde{2}0$ 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 1, which lends credence to there possibly being some leakage.

E 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 4.6 as Theorem 4.5 proof's relies on it, it being the weaker of the two.

Theorem 4.6 (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 4.5 (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 4.6. 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.