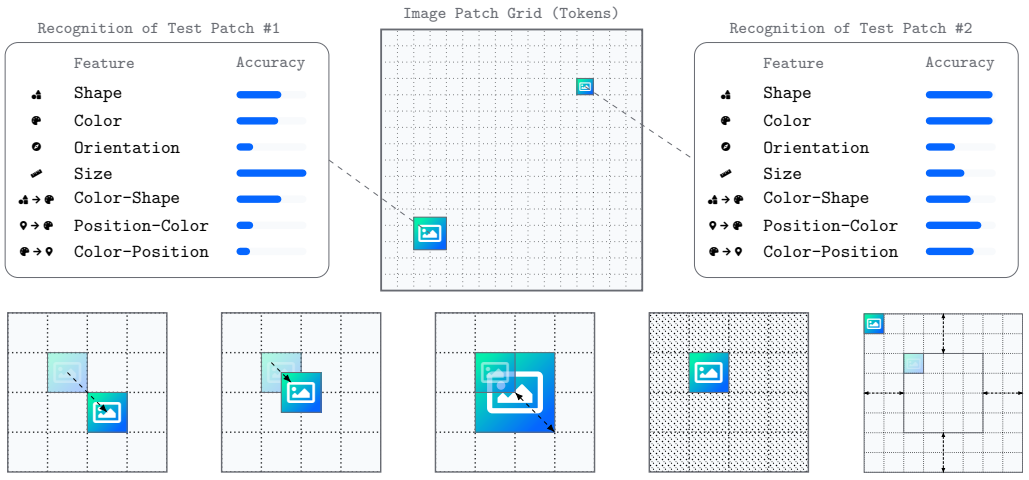
Same same, but different: VLMs Do Not Recognize Features Equally Anywhere in an Image

Anonymous authors
Paper under double-blind review

Abstract

Vision-Language Models (VLMs) have achieved impressive performance on a wide range of tasks requiring visual understanding. Despite this progress, their predictions remain surprisingly brittle and sensitive to small spatial variations. Using synthetic experiments that isolate specific visual factors, we *systematically* study *anisotropic feature recognition* in VLMs. We find that modern VLMs reliably recognize simple atomic features contained within single patches, but continue to struggle with more complex atomic and compositional features. Crucially, the same feature can be recognized with considerably different fidelity depending on its position in the input image, with spatial anisotropy patterns that vary across feature types and models. We further show that fragmentation across patches is rarely harmful, while upsampling a feature over more patches substantially improves recognition, particularly for compositional features. Conversely, visual distractors and longer vision-token sequences degrade feature recognition and introduce additional positional biases. Taken together, our results reveal that feature recognition in VLMs remains strongly anisotropic and sensitive to the allocation of visual information across tokens.



(a) on-grid (b) fragmentation (c) upsampling (d) distractors (e) sequence length

Figure 1: **Anisotropic feature recognition.** We test how the recognition of individual features encoded in small test patches is affected by (a) translation on the patch grid, (b) translation that results in fragmentation over multiple patches, (c) upsampling, (d) surrounding distractors, and (e) a placement in longer token sequences.

1 Introduction

Vision-Language Models (VLMs) have made incredible leaps on tasks with impressive complexity, such as visual question answering (Antol et al., 2015; Yue et al., 2024), autonomous driving (Sima

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.

et al., 2025), and abstract reasoning (Chollet, 2019). Yet, at the same time, their performance can rapidly drop under slight variations (Nezhurina et al., 2025; Vo et al., 2026; Zheng et al., 2024; Zong et al., 2024), suggesting that models overfit to shortcuts that are sufficient to solve in-distribution examples but do not develop true generalization (Golchin & Surdeanu, 2024; Li & Flanigan, 2024; Wu et al., 2024), which would exclusively depend on robust core features (Geirhos et al., 2020). This issue is especially acute for compositional generalization, where scaling alone may not cover the combinatorial space of feature combinations (Uselis et al., 2025).

A notable weakness in current Vision-Language Models (VLMs) is their lack of geometric robustness, specifically regarding translational variance. Research indicates that these models often yield different answers depending on where features are positioned, failing to maintain the spatial consistency expected of sophisticated visual systems (Rosenfeld et al., 2025). In this paper, we are concerned with translation (in)variance with respect to the generated response, not the latent representation. For instance, when a model is asked to recognize the animal in an image, it should recognize the presented animal, independent of where in the scene it is located. The latent representation, on the other hand, may be equivariant, as the model will additionally encode the position information at some stage.

Commonly, prior work attributes translation variance in Vision Transformer (ViT)-based models to the discretization of images into non-overlapping tokens (Ding et al., 2023; Rojas-Gomez et al., 2024), resulting in the fragmentation of features over multiple tokens. Prior work has largely studied this by translating natural images (Shifman & Weiss, 2024; Rosenfeld et al., 2025). However, answering questions on these images often requires a compositional understanding of multiple features, including binding attributes to the correct visual entities (Koishigarina et al., 2026; Campbell et al., 2024), each of which may be (partially) scattered over one or multiple input tokens. Translating an image impacts all of these tokens simultaneously, complicating a clean isolation of the underlying root cause.

In this study, we take a step back and switch to a clean and systematic approach to understand *anisotropic feature recognition*[1] (Fig. 1). We construct synthetic visual stimuli probing different atomic and compositional features, where only one relevant patch in the image is active, *i.e.*, it fully contains the feature relevant for the query. This patch is positioned such that it is initially represented by exactly one token. All other tokens are considered passive background tokens, irrelevant for the query. We show that a simple translation over the grid without fragmenting the feature already strongly affects the prediction, and the resulting spatial variance depends on both the feature type and model. We further test the influence of fragmentation caused by translation, of using larger feature patches that span multiple tokens, of distractors, and of the sequence length (resolution). Taken together, our study cleanly isolates changes on different tested visual factors as shown in Figs. 1a to 1e.

We summarize our contributions as follows:

- We propose a testbed of synthetic stimuli to evaluate the response shift on atomic and composite features of VLMs based on the spatial location of the presented feature (§3).
- We systematically evaluate feature recognition on three open VLMs across five axes of visual factors: on-grid position, fragmentation, scale, visual distractors (noise and clutter), and the sequence length of the vision tokens–showing that spatial anisotropy persists even in recent models (§4). While model progress has largely improved single-patch recognition of simple atomic features, complex or compositional features remain lossy, unevenly represented across the patch grid (§4.1); fragmentation rarely decreases performance (§4.2), upsampling often helps (§4.3), while distractors (§4.4.1) and longer visual contexts (§4.4.2) reliably degrade performance.

2 Preliminaries and Background

VLMs typically integrate ViT-based encoders and large language models (LLMs) by projecting vision-derived tokens into the language space (Tsimpoukelli et al., 2021; Chen et al., 2022; Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023; Zhu et al., 2024; Zhang et al., 2024). While ViTs

[1]Code for review: https://anonymous.4open.science/r/anisotropic-features



Figure 2: Representative examples from the atomic and compositional feature families.

offer high learning capacity, their architectural design introduces specific biases. In this section, we formalize key components and issues of the ViT pipeline related to our study.

Image tokenization and patching. Let an input image be denoted by $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and channel depth, respectively. Images are then usually discretized into a grid of $N \times M$ non-overlapping patches, each of size $P \times P$. The resulting sequence length is defined by $L = NM$, where $N = H/P$ and $M = W/P$ (Dosovitskiy et al., 2021). Each patch is flattened into a vector and mapped to a latent space via a learnable linear projection. It is well-established that ViT performance inversely scales with patch size P; smaller values of P generally yield higher predictive power as the model captures finer-grained spatial details (Dosovitskiy et al., 2021; Wang et al., 2025a), suggesting that standard patch representations with $P > 1$ act as inherently lossy compressions of local features, particularly evident if the corresponding image patch is polysemantic (Chen et al., 2024a). Furthermore, ViTs require strong training to compensate for their lack of inductive bias (Touvron et al., 2021). Thus, spatial feature robustness is not inherent and may be learned unevenly across patches. Indeed, multiple papers have shown that not all tokens are equally utilized: Xiao et al. (2024) discovered attention sinks, where representations collapse into a few high-norm tokens. Darcet et al. (2024) demonstrated that tokens corresponding to background patches may be repurposed as registers, and Kim et al. (2023) showed a Gaussian-like attention bias in ViT classifiers, where outer patches are deemed less important for generations. Orthogonally, our work studies how well certain features can be decoded from single patches. We find significant fluctuations depending on the model, position, and the kind of feature across the patch grid. The emergent spatial variance patterns often do not follow a Gaussian distribution and instead highlight clusters of "lossiness".

Shift-variance and spatial biases. Recent empirical studies suggest that architectural choices in ViT, like discretization and positional embeddings, are primary drivers of performance degradation under spatial perturbation (Wang et al., 2025a; Rojas-Gomez et al., 2024; Ding et al., 2023; Chu et al., 2023), a rampant issue that scale alone does not seem to solve (Shifman & Weiss, 2024). Compared to prior work, which focuses on studying and improving shift-equivariance or -invariance for entire input images, we take a step back and focus on the systematic evaluation of existing anisotropic biases inherent in well-adopted VLMs. Our findings show that discretization alone is rarely an issue; instead, a much larger driver of variance is the fluctuation of feature fidelity across the patch grid.

Needle-in-a-haystack. Our work is related to needle-in-a-haystack tests, where a needle is embedded in a haystack of irrelevant tokens. Often this methodology is applied to study the retrieval of information in increasingly longer sequences (Kamradt, 2023; Liu et al., 2024b) or for increased input resolution for input images (Li et al., 2026; Chen et al., 2025; Lu et al., 2025). Prior works have already established that the recognition of the needle is increasingly more challenging with increased haystack size in text or multi-modal context (sequences of images) and have determined various trends (Kamradt, 2023; Liu et al., 2024b; Levy et al., 2024; Wu et al., 2025; Tian et al., 2025). We extend their experiments to single images where we vary the needle (feature, position, size) and haystack (distractors) primarily *without* increasing the haystack (except for §4.4.2). Importantly, our work pays significant attention to the features that the needle encodes.

3 METHODOLOGY

Our methodology revolves around a diagnostic dataset, loosely inspired by dSprites (Matthey et al., 2017): we create small feature patches of size $F \times F, s.t., F \leq \min(W, H)$ (typically, $F = P$, where P is model-specific) and translate them to various positions of a larger canvas image (typically a uniform black background), while controlling all other external factors and measuring the recognition rate of the contained feature(s). However, complementary to dSprites, we are interested in VLM predictions through visual question answering (VQA) (Antol et al., 2015; Zhang et al., 2016; Goyal et al., 2017) and not the representation in the latent space. We also extend our testing to compositional features. The following subsections explain our methodology in more detail.

Features. We create stimuli that test four *atomic* and three *compositional* features (specifically, *attribute binding*) with examples shown in Fig. 2. Each stimulus is accompanied by one or more prompts documented in §F, probing for the respective feature(s). Atomic features represent inherent, standalone properties of a single entity, and we test the following:

- **Shape** (66 samples): All English letters and digits and six simple filled geometric shapes and their outline variants rendered in white (in three different line widths).
- **Color** (6 samples): Patches filled with one of six colors.
- **Orientation** (144 samples): Filled arrows in six different colors and three different widths pointing in all directions in $45°$ steps. The task is to recognize the cardinal direction.
- **Size** (54 samples): Filled rectangle patches across nine aspect-ratio variants in six different colors. The task is to recognize the aspect-ratio (wide, tall, square).

Compositional features require the ability to correctly link multiple atomic features to their specific objects, parts, or spatial relationships, and we test:

- **Shape-Color** (150 samples): We probe for the color of a filled shape (excluding squares) on a differently colored background.
- **Position-Color** ($2 \cdot 60$ samples): We generate squares, split vertically or horizontally, where each half is filled with a different color, and independently probe the color of *each* half by position.
- **Color-Position** ($2 \cdot 60$ samples): This is the inverse task of *Position-Color*. We probe for the position of *each* half by color.

These features are meant to be a useful subset of features present in real-world tasks but are naturally limited. We show that our findings are not limited to these synthetic stimuli in §A.

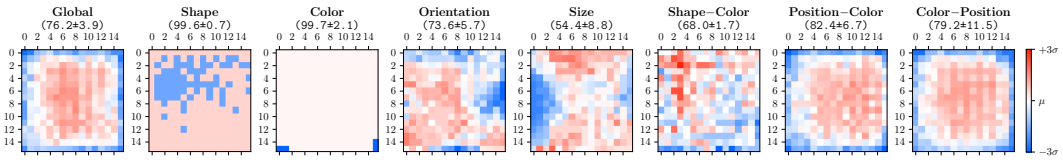
Evaluation. We execute most experiments with three different open VLMs: `Qwen-3.6-35B-A3B` (Qwen Team, 2026c) and `InternVL3.5-30B-A3B` (Wang et al., 2025b) represent modern Mixture-of-Expert (Shazeer et al., 2017) models with high-resolution support; `LLaVA1.5-7B` (Liu et al., 2024a) is a highly popular auto-regressive VLM and was one of the first successful open-source VLMs, albeit that its capabilities are quite dated by today. `InternVL` rescales and partitions inputs into tiles matching the encoder's native resolution of 448^2px. These tiles are encoded independently alongside a downsampled global thumbnail (Chen et al., 2024b; Wang et al., 2025b). In contrast, `Qwen` processes inputs at their native spatial resolution (with a minimum of 256^2px) (Qwen Team, 2026a;b;c). `LLaVA1.5` is limited to process images of 336^2px. Together, these models signify the rapid evolution of models over the last three years.

We generate answers by passing the stimulus image with a corresponding multiple-choice prompt in a single-turn assistant completion. We process the outputs via light heuristics (case-sensitivity, whitespace removal, mapping of hallucinated option keys to the actual answer) and string-match them against ground truths. All completions are generated under greedy decoding, without a system prompt, and without thinking/reasoning. For a discussion of language priors, please see §C.

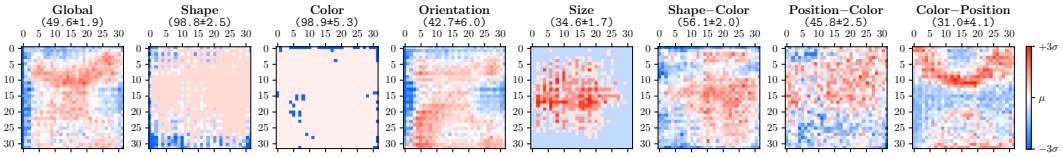
Measuring spatial performance. Our primary metric is accuracy $acc(f, i, j)$ of feature f at position (i, j). To obtain an aggregated measurement of the variance in spatial feature fidelity over a position grid, we additionally compute a grid-size-normalized, isotropic total variation (Rudin et al., 1992):

$$TV = \frac{1}{\sqrt{NM}} \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} \sqrt{\big(acc(f, i+1, j) - acc(f, i, j)\big)^2 + \big(acc(f, i, j+1) - acc(f, i, j)\big)^2}.$$

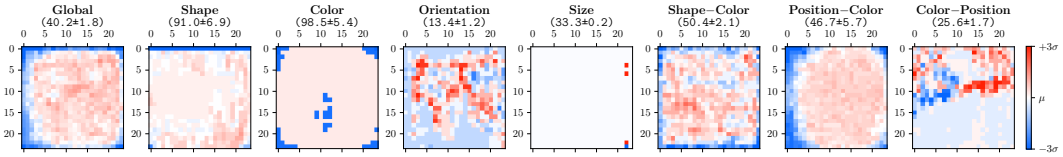
Intuitively, we treat the accuracy grid as a 2D image: low total variation means neighboring positions have similar feature fidelity, while high total variation indicates sharp spatial changes. To compensate for the imbalanced number of samples across tasks, we perform a negated permutation test to compare task performances (Wooldridge, 2013): $Z_{TV} = -(TV - \mu)/\sigma$, where μ and σ denote the mean and standard deviation of TV, respectively, obtained from 1,000 trials using spatially permuted measurements. We verified that the distribution of TV is approximately Gaussian in our use cases, confirming the validity of this approach. In this standardized metric, negative values for Z_{TV} indicate noisy spatial variance patterns, while positive values indicate clustered variance patterns. A value of 0 signifies no more spatial structure than a random permutation.



(a) `Qwen3.6-35B-A3B` (Qwen Team, 2026c)



(b) `InternVL3.5-30B-A3B` (Wang et al., 2025b)



(c) `LLaVA1.5-7B` (Liu et al., 2024a)

Figure 3: **Spatial distribution of feature fidelity.** We compute the accuracy for each feature for all positions in the patch grid of minimum supported resolution for different models. Heatmaps show the positive (red) and negative deviation (blue) for $\pm 3 std$ from the average performance over all patches. The number below the feature name indicates the spatial $mean \pm std$ accuracy.

4 Experimental Results

Our experiments separate the factors behind spatial brittleness that are usually mixed together when images are translated as a whole. We begin with the cleanest case: a single relevant feature is placed exactly on the patch grid, so that it is fully contained in one vision token. Even then, feature fidelity is strongly position-dependent, showing that spatial anisotropy is not just an artifact of off-grid fragmentation (§4.1). This motivates testing fragmentation directly: splitting the same feature across neighboring tokens is rarely harmful and can even slightly improve recognition (§4.2). Since fragmentation does not increase the amount of visual evidence, we then upsample the feature across multiple dedicated tokens, which substantially improves recognition, especially for compositional features (§4.3). Finally, we test whether these effects persist beyond isolated patches, finding that irrelevant distractors and longer vision-token sequences reduce feature fidelity and alter spatial bias patterns (§4.4).

4.1 Feature Fidelity is Spatially Distorted

Our main claim is that feature fidelity is not equally distributed across the patch grid. We verify this by applying our methodology described in §3, placing feature patches at on-grid positions $\forall (i, j) \in \{1, \dots, N\} \times \{1, \dots, M\}$ and testing the accuracy. For each model, we place a feature patch of size $F \times F$ with $F = P$, and utilize an image size that corresponds to the smallest supported sequence that doesn't cause resizing. This equals to 256^2px for `Qwen`, 448^2px for `InternVL`, and 336^2px for `LLaVA` models.

Compression of features into one patch. To understand which features can be recovered from a single patch, we compare the average performance of each feature over all grid positions to a feature rendered at full image resolution (see Tab. 1 for a results table). Across all evaluated models, atomic features exhibit less recognition degradation than compositional features when compressed to a single patch. However, while the compression minimally impacts easy atomic features like shape and color, we observe a significant degradation for the more complex atomic features of orientation and size. For compositional features, performance degrades in our on-grid setting, yet remains consistently above chance despite compression, suggesting that feature binding is lossy rather than

Table 1: **Mean performance per feature.** We show the accuracy on the feature rendered at full resolution of the image [full-res.] and as a spatial average (mean ± std) over all on-grid positions, where the feature is exactly as large as one patch.

Model	Type	Accuracy per Feature [%]						
		Atomic				Compositional		
		Shape	Color	Orientation	Size	Shape-Color	Position-Color	Color-Position
Chance		10.6	16.7	12.5	33.3	16.7	16.7	25.0
`Qwen3.6-35B-A3B` (Qwen Team, 2026c)	full-res.	98.5	100.0	90.3	98.1	99.3	100.0	100.0
	on-grid	99.6±0.7	99.7±2.1	73.6±5.7	54.4±8.8	68.0±1.7	82.4±6.7	79.2 ± 11.5
`InternVL3.5-30B-A3B` (Wang et al., 2025b)	full-res.	62.1	100.0	54.2	57.4	100.0	100.0	100.0
	on-grid	98.8±2.5	98.9±5.3	42.7±6.0	34.6±1.7	56.1±2.0	45.8±2.5	31.0 ± 4.1
`LLaVA1.5-7B` (Liu et al., 2024a)	full-res.	92.4	100.0	9.0	33.3	86.7	96.7	50.0
	on-grid	91.0±6.9	98.5±5.4	13.4±1.2	33.3±0.2	50.4±2.1	46.7±5.7	25.6 ± 1.7

fully absent. Furthermore, we identify asymmetric binding effects: color-position binding exhibits more pronounced degradation than its inverse. Given that all models achieve a fairly high accuracy on the binding tasks in full resolution, our results suggest that binding benefits from distributing features across multiple vision tokens. This hypothesis is further supported by experiments in §4.3, where upscaling a feature patch over additional tokens yields substantial improvements in compositional feature recognition.

Summarizing, our evaluation demonstrates that VLMs can recognize simple features from a single patch, but struggle with more complex features and binding. While there is a strong improvement in the tested models with chronological order (and also their total number of parameters), we still see a notable reduction in average performance for these complex features–importantly, we also see a strong variance depending on the patch position that we will discuss next.

Spatial variance of feature fidelity. Turning to spatial variance, we observe in Fig. 3 that not a single feature is isotropic. While spatial variance patterns across all samples are often, but not always, Gaussian-distributed (similar to the findings in (Kim et al., 2023)), they differ by feature and model. In isolation, they rarely follow a Gaussian-like distribution. For instance, while `LLaVA` struggles to recognize shapes within the first row of patches (Fig. 3c), it performs reliably at other positions, and color recognition is notably weakest in the center. The spatial pattern can also be different for inverse tasks (*cf*., position-color *vs*. color-position in Figs. 3b and 3c, right columns). Interestingly, spatial bias can "revive" certain features: *e.g*., while `LLaVA` has a near-random performance on average of 25.6% for color-position, numerous positions in the upper half of the grid can boost the accuracy up to 32.2%. Similarly, certain positions can "kill" a feature: `Qwen`'s recognition of size drops from an average of 54.4% and a max of 72.2% to random chance at certain positions near the left edge (Fig. 3a). Spatial bias can (but does not have to) emerge even if the spatially averaged accuracy is near random (*cf*. the negligible fluctuation for size in `LLaVA`). Overall, we see the least average anisotropy in `Qwen` and the highest in `InternVL` (Fig. 4). We attribute `InternVL`'s higher anisotropy to its larger number of vision tokens. We will show that the number of vision tokens strongly affects spatial bias in §4.4.2. Generally, we expect anisotropy to improve in future models.

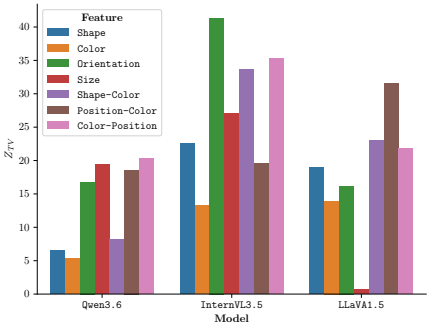


Figure 4: **Spatial variation.**

4.2 Patch Fidelity vs. Fragmentation

Previous studies on translational variance often measured performance under the translation of the entire scene. However, the answers will often depend on multiple features, each potentially being distributed over multiple patches: This confounds two potential issues: different patches may encode features with different fidelity (which we have shown in §4.1), and fragmentation, *i.e*., the translation causes the feature to be (partially) straddled over multiple patch tokens. In this section, we leave our on-grid setting and place feature patches such that they are equally fragmented over two and four

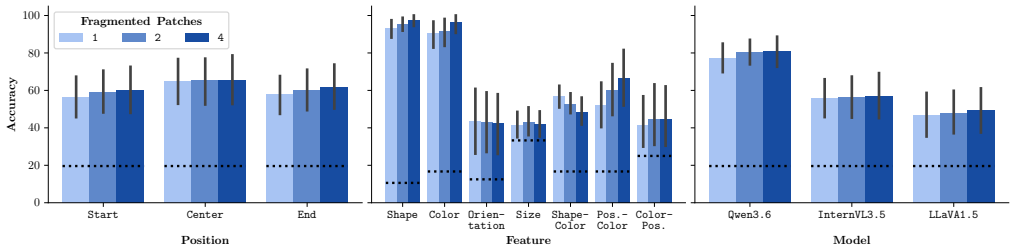


Figure 5: **Feature recognition under fragmentation.** We aggregate over positions, features, and models. The error bars denote the 95% CI. Dotted horizontal lines show the random chance.

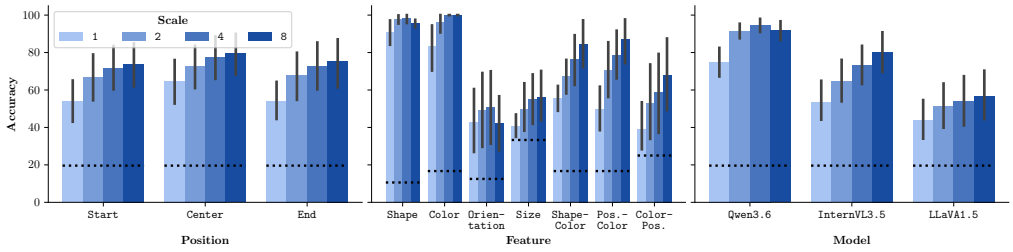


Figure 6: **Feature recognition under upsampling.** We aggregate over positions, features, and models. The error bars denote the 95% CI. Dotted horizontal lines show the random chance.

tokens, then compare them to the on-grid results at adjacent positions. The feature patch size remains unchanged. We test positions in the center and in the top-left and bottom-right corners of 2×2 grids.

Surprisingly, the aggregated results in Fig. 5 show small improvements in performance when features are fragmented over multiple patches, contrasting previous findings that call for merging (Renggli et al., 2022). Generally, we see greater improvements at the beginning and end of the image patch sequence and only minor changes in the center of the sequence (Fig. 5, left). However, these trends are not universal: we find that the compositional shape-color task performance consistently decreases under fragmentation (Fig. 5, middle). Features may also show different trends depending on the position, *e.g.*, we find that the atomic shape task benefits from fragmentation when the active patch is placed at the beginning or end of the image sequence but suffers from fragmentation in the center on `LLaVA` as depicted in the detailed evaluation plots in Fig. 14 (third row, first column). The opposite trend can be observed on the compositional color-position task and `InternVL` (Fig. 14, second row, last column). Finally, different models can show different trends for the same feature, *e.g.*, the orientation task benefits from fragmentation on `LLaVA` and `InternVL`, but suffers on `Qwen` (Fig. 14, third column).

4.3 Feature Recognition Improves with Upsampling

Following the experiments in §4.2, we now look at the effects of resizing specifically *upsampling*– which also fragments features over multiple patches but, unlike the previous test, ensures that the corresponding tokens are exclusively populated by our feature patch. We perform this experiment by *upsampling* the original feature patches to $F = sP$, where $s \in \{1, 2, 4, 8\}$ and place it in the center and in the top-left and bottom-right corners of the image. We upsample using nearest neighbors, which does not inject any new information relative to the original patch, and avoids aliasing, which has been linked to robustness issues (Zhang, 2019; Azulay & Weiss, 2019; Grabinski et al., 2022).

The results in Fig. 6 demonstrate that performance on all models and all feature tasks generally significantly increases with larger s, with the largest improvements from s=1 to s=2. The effects are much stronger than under comparable fragmentation (as in §4.2), suggesting that a stronger utilization in terms of the area of a patch improves the fragmented features' representation. We see particularly strong improvements for the compositional features (Fig. 6, middle). One notable outlier is the orientation task (Fig. 6, middle), where accuracy initially improves, then eventually starts declining for larger s. Interestingly, we also find a few individual outliers at the max scale, where performance

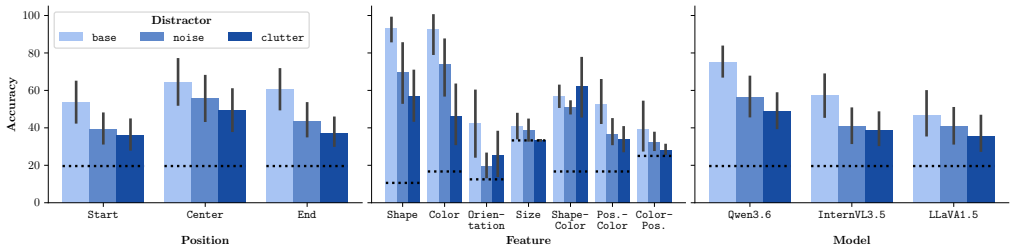


Figure 7: **Feature recognition under visual distraction.** We aggregate over positions, features, and models. The error bars denote the 95% CI. Dotted horizontal lines show the random chance.

can start to decrease, suggesting that upsampling may follow an inverted U-trend. This is consistent with one of the full-resolution feature observations in §4.1, where `InternVL`'s shape performance dropped to 62.1% when features were generated (not upsampled) at the full image resolution, but is otherwise near perfect. Trends are fairly similar for all tested models, although gains in `Qwen` seem to saturate or even decrease with larger scale (Fig. 6, right). `InternVL` also shows a steeper improvement than `LLaVA`. For non-aggregated plots, we refer to Fig. 15.

4.4 Feature Recognition Depends on the Context

The previous experiments have been conducted by placing feature patches on a neutral black background in images of the minimum supported resolution. In real-world tasks, relevant features will often be embedded in a more complex background (*context*), which, even if logically unrelated, may impact the prediction (*e.g.*, see (Sagawa et al., 2020) for an example on ImageNet (Deng et al., 2009)). Consequently, we test two dimensions of changes to the context via visual distractions (§4.4.1), and by embedding the feature in a large token sequence (§4.4.2).

4.4.1 Feature Recognition Deteriorates under Visual Distraction

We test two kinds of distractors at extreme ends of entropy: noise and clutter. Noise embeds the active patch in a meaningless, high-entropy pattern, yet may cause non-robust encoders to hallucinate non-existent features. We realize this by replacing the black background with uniform color noise. The other end of our distractors uses clutter, surrounding the active patch at (i, j) by lower-entropy patches with a clear semantic meaning for humans (Whitney & Levi, 2011) and may decrease the localization of the patch, especially under shared attributes (Campbell et al., 2024). We realize this by populating all patches $\forall(i', j') \in \{1, \dots, N\} \times \{1, \dots, M\}, s.t., (i', j') \neq (i, j)$ with features unrelated to the prompt (see Tab. 6 for a list of those). Example stimuli for both distractors can be found in Fig. 13.

Unsurprisingly, the results in Fig. 7 clearly demonstrate a degradation of performance under distraction, where clutter causes the most degradation. Trends in different positions are fairly similar for the most part, albeit with typically stronger magnitude at the start and end of the patch sequence (Fig. 7, left). For instance, `Qwen`'s shape recognition seems to be mostly affected by visual clutter if the active patch is at the beginning of the sequence (Fig. 16, first row, first column). However, we also find instances where clutter *improves* performance: compositional shape-color on `Qwen` and `InternVL`, but not on `LLaVA` (Fig. 16, fifth column). On `LLaVA`, we observe some minor improvements on this task compared to the baseline under noise, and only if the active patch is placed at the start or center of the patch sequence (Fig. 16, third row, fifth column). For the most part, trends are stable across positions; one notable outlier is the size feature on `Qwen`, where performance slightly improves with noise on start and center positions, but almost as strongly deteriorates due to clutter if the active patch is placed at the end of the sequence (Fig. 16, first row, fourth column).

4.4.2 Feature recognition Unevenly Deteriorates with Sequence Length

In this final experiment, we extend the resolution, effectively embedding the feature patch in a longer vision token sequence. We use a resolution of 4928^2px for `InternVL` corresponding to a token increase of $121\times$ from $L{=}32^2$ to $L{=}352^2$. We apply the same rate of increase for `Qwen`, resulting in

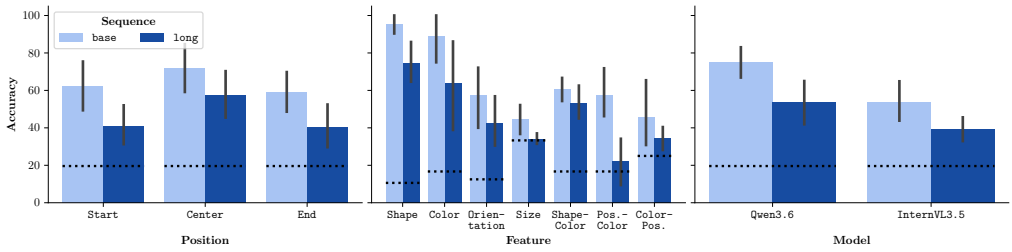


Figure 8: **Feature recognition under long context.** We aggregate over positions, features, and models. The error bars denote the 95% CI. Dotted horizontal lines show the random chance.

a resolution of 2816^2px, increasing $L{=}16^2$ to $L{=}176^2$. We do not include `LLaVA` in this experiment, due to its lack of support for higher resolutions.

The results in Fig. 8 clearly show a decrease in performance with an increase in sequence length, aligning with prior findings (Li et al., 2026; Chen et al., 2025; Lu et al., 2025), occasionally dropping to random chance. The increased sequence length strongly affects active patches at the start or end of the patch sequence, but also reduces performance in the middle. All atomic and compositional features are negatively affected through the increased sequence length, and we do not observe clear trends in atomic *vs*. compositional features. Instead, we observe large differences within these groups: compositional position-color sees the strongest degradation, while shape-color is the feature affected the least. Relative to their original performance, both `InternVL` and `Qwen` degrade equally to approximately 72% of the original performance. We provide the non-aggregated analysis in Fig. 17 and estimate the full variance pattern in §B.

5 Conclusion

We studied how reliably open VLMs recover visual features when the semantic content of the image is held fixed but the patch-level allocation of that content changes–a setting that is only realizable with synthetic stimuli. Our experiments show that translation variance in VLMs is, contrary to popular belief, not a consequence of fragmentation: even when a feature is fully contained in a single vision token, its recognition can vary substantially across the patch grid. These anisotropic feature representations are feature- and model-dependent and can make particular positions unusually favorable or unfavorable–a trend that remains visible in recent models. We find that fragmentation often improves recognition, even more strongly if the feature is upscaled before the fragmentation. Increasing visual context, either through additional content (independent of distractor entropy) or a longer token sequence induced by higher resolution, further amplifies anisotropy. Taken together, these results suggest that robust VLM perception requires more than higher benchmark accuracy or larger input resolutions. Models should preserve task-relevant visual features uniformly across patch positions, maintain their binding under spatial compression, and resist interference from irrelevant visual context. Our anisotropic feature recognition testbed provides a simple diagnostic for these properties and a way to separate failures due to patch location, fragmentation, scale, distractors, and sequence length.

Limitations. To effectively isolate patch-level effects, our study intentionally utilizes simple synthetic stimuli and VQA-style multiple-choice prompts. Consequently, this approach does not reflect the full diversity of natural scenes where multiple confounding effects may coexist. However, we show similar findings on exemplary photorealistic tasks in §A. Furthermore, our curated set of atomic and compositional features is inherently finite and may underrepresent other visual characteristics. However, we expect the selected features to be a reasonable subset of features in real-world tasks. Additionally, to manage computational costs, our experiments outside the on-grid setting probe three representative sequence locations rather than exhaustively testing all possible patch positions or combinations of perturbation axes. This sampling strategy means our findings in these areas may over- or understate certain effects (we give some insights for sequence length experiments in §B). Finally, because our analyses rely on generated answers rather than latent probes, our measurements evaluate end-to-end feature utilization in the final response, but do not establish whether features are absent or not utilized.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding and to polish the language in the manuscript. We have not used generative AI tools for all other tasks with required disclosure. We have reviewed all AI-assisted work. Specifically, generated code was verified and tested for correctness by the authors. Polished sections were proofread by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We refer the reader to §D for details about reproducibility. We commit to fully open-source the code and raw responses to all stimuli upon acceptance. A snapshot for review can be found here: https://anonymous.4open.science/r/anisotropic-features.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research (JMLR)*, 20(184), 2019. URL http://jmlr.org/papers/v20/19-519.html.

Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=Q5RYn6jagC.

Delong Chen, Samuel Cahyawijaya, Jianfeng Liu, Baoyuan Wang, and Pascale Fung. Subobject-level image tokenization. *arXiv preprint arXiv:2402.14327*, 2024a.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. doi: 10.1109/CVPR52688.2022.01750.

Liangyu Chen, James Burgess, Jeffrey J. Nirschl, Orr Zohar, and Serena Yeung-Levy. The Impact of Image Resolution on Biomedical Multimodal Large Language Models. In *MLHC*, Proceedings of Machine Learning Research, 2025.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12), 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4231-5. URL https://doi.org/10.1007/s11432-024-4231-5.

François Chollet. On the Measure of Intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional Positional Encodings for Vision Transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=3KWnuT-R1bh.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. doi: 10.1109/CVPR.2009.5206848.

Peijian Ding, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. Reviving Shift Equivariance in Vision Transformers. *arXiv preprint arXiv:2306.07470*, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Shahriar Golchin and Mihai Surdeanu. Time Travel in LLMs: Tracing Data Contamination in Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Julia Grabinski, Janis Keuper, and Margret Keuper. Aliasing and adversarial robust generalization of CNNs. *Machine Learning*, 111(11), 2022. ISSN 1573-0565. doi: 10.1007/s10994-022-06222-8. URL https://doi.org/10.1007/s10994-022-06222-8.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary Position Embedding for Vision Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, 2024.

Gregory Kamradt. LLMTest_NeedleInAHaystack. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023. [Online; accessed 1. May 2026].

Bum Jun Kim, Hyeyeon Choi, Hyeonah Jang, and Sang Woo Kim. Understanding Gaussian Attention Bias of Vision Transformers Using Effective Receptive Fields. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2023. URL https://papers.bmvc2023.org/0214.pdf.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP Behaves like a Bag-of-Words Model Cross-modally but not Uni-modally. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=DldwXCCP25.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistic (ACL)*, 2024. doi: 10.18653/v1/2024.acl-long.818. URL https://aclanthology.org/2024.acl-long.818/.

Changmao Li and Jeffrey Flanigan. Task contamination: language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, AAAI'24/IAAI'24/EAAI'24, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i16.29808. URL https://doi.org/10.1609/aaai.v38i16.29808.

Chenxu Li, Zhicai Wang, Yuan Sheng, Xingyu Zhu, Yanbin Hao, and Xiang Wang. Res-Bench: Benchmarking the Robustness of Multimodal Large Language Models to Dynamic Resolution Input. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a. doi: 10.1109/CVPR52733.2024.02484.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Dongchen Lu, Yuyao Sun, Zilu Zhang, Leping Huang, Jianliang Zeng, Mao Shu, and Huo Cao. Internvl-x: Advancing and accelerating internvl series with efficient visual token compression. *arXiv preprint arXiv:2503.21307*, 2025.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. *arXiv preprint arXiv:2406.02061*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. 2019.

Qwen Team. Qwen3.5: Accelerating Productivity with Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5, 2026a. [Online; accessed 1. May 2026].

Qwen Team. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model. https://qwen.ai/blog?id=qwen3.6-27b, 2026b. [Online; accessed 1. May 2026].

Qwen Team. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All. https://qwen.ai/blog?id=qwen3.6-35b-a3b, 2026c. [Online; accessed 1. May 2026].

Cedric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. Learning to Merge Tokens in Vision Transformers. *arXiv preprint arXiv:2202.12015*, 2022.

Renan A. Rojas-Gomez, Teck-Yian Lim, Minh N. Do, and Raymond A. Yeh. Making Vision Transformers Truly Shift-Equivariant. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Amir Rosenfeld, Neta Glazer, and Ethan Fetaya. Questioning the Stability of Visual Question Answering. *arXiv preprint arXiv:2511.11206*, 2025.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1), 1992. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(92)90242-F. URL https://www.sciencedirect.com/science/article/pii/016727899290242F.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Ofir Shifman and Yair Weiss. Lost in Translation: Modern Neural Networks Still Struggle With Small Realistic Image Transformations. *arXiv preprint arXiv:2404.07153*, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, 2025.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and Mitigating Position Bias of Multi-image Vision-Language Models. *arXiv preprint arXiv:2503.13792*, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal Few-Shot Learning with Frozen Language Models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/01b7575c38dac42f3cfb7d500438b875-Paper.pdf.

Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Does Data Scaling Lead to Visual Compositional Generalization? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 2025.

Guido van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), 1995.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision Language Models are Biased. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=DG4S2OlGQA.

Feng Wang, Yaodong Yu, Wei Shao, Yuyin Zhou, Alan L. Yuille, and Cihang Xie. Scaling Laws in Patchification: An Image Is Worth 50,176 Tokens And More. In *Proceedings of the International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu,

Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

David Whitney and Dennis M. Levi. Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 15(4), 2011. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2011.02.005. URL https://www.sciencedirect.com/science/article/pii/S1364661311000325.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. 5 edition, 2013.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David Chan. Visual Haystacks: A Vision-Centric Needle-In-A-Haystack Benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=9JCNPFL1f9.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. doi: 10.18653/v1/2024.naacl-long.102. URL https://aclanthology.org/2024.naacl-long.102/.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and Answering Binary Visual Questions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=d4UiXAHN2W.

Richard Zhang. Making Convolutional Networks Shift-Invariant Again. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/zhang19a.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large Language Models Are Not Robust Multiple Choice Selectors. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool Your (Vision and) Language Model With Embarrassingly Simple Permutations. *arXiv preprint arXiv:2310.01651*, 2024.

Overview of the Appendix

A Towards "Real-World" Examples **16**

B Comparison of Spatial Variance Patterns under Long(er) Context **17**

C Language Priors **17**

C.1 Option Order Sensitivity 18

C.2 Prompt Sensitivity 18

C.3 Response Parsing 20

D Reproducibility Statement **20**

E Visualization of Test Stimuli **21**

F List of Prompts per Feature **21**

G Full Evaluation Results **22**

A Towards "Real-World" Examples

While our experiments in §4 relied on synthetic features to disentangle different visual factors, we want to demonstrate that anisotropy also affects photorealistic data, more closely resembling real-world data. We create two simple settings to test this:

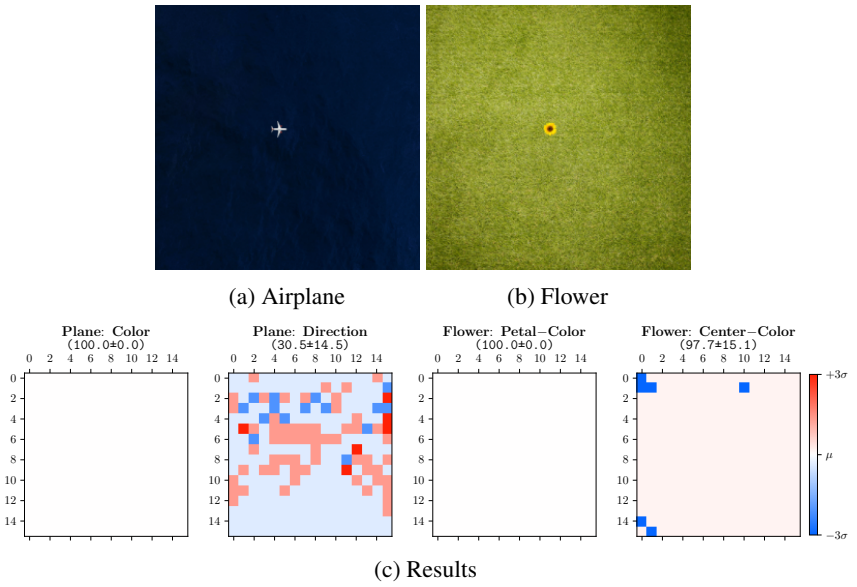
- **Airplane:** we cut out and downsample a bird's-eye-view image of a commercial airplane to P^2px, rotate it in four different orientations (flying up, down, left, right), and place it over a background of the deep sea. We probe two recognition tasks: the color of the plane (*white*) and the direction it is flying.
- **Flower:** we cut out and downsample an image of a sunflower to P^2px, and place it over a background of grass. The task is to recognize the petal color and center color of the flower (*yellow, black*, respectively).

See Figs. 9a and 9b for examples of the resulting stimuli.[2]

We then repeat the experiments in §4.1, by placing the feature patches (airplane or flower) over all possible grid positions on the corresponding background. We limit experiments to `Qwen3.6-35B-A3B` (Qwen Team, 2026c), as the patches are poorly recognizable at $P < 16$. Note that, unlike the clean methodology in §3, the background under the active patch now shows minor differences in pixel space but is still stable in semantic space.

The results in Fig. 9c show no spatial bias and perfect performance for the plane and petal color tasks, but some anisotropy for the other tasks, confirming that our findings extend to real-world settings. `Qwen` struggles to disentangle the center from the petal colors, and will respond with the petal color when asked about the center color in some locations. It also struggles heavily to detect the plane's orientation, responding with "right" to 64% of the samples. This suggests yet another bias in VLMs (Vo et al., 2026), and indeed an analysis of a reasoning trace (single sample with thinking mode enabled) shows that `Qwen` believes the active patch to show a plane emoji, which it claims to be typically oriented right, or up ("up" is the second most common response at 22.9%). In any case, the spatial anisotropy is exceptionally high for this task.

[2]Asset credits: Image by Iftikhar Alam on Vecteezy [Vecteezy License]; Photo by Cath Smith on Unsplash [Unsplash License]; Photo by Gabor Kozmon on Unsplash [Unsplash License]; Photo by Akshay Chauhan on Unsplash [Unsplash License]



(a) Airplane (b) Flower

(c) Results

Figure 9: **Anisotropy on photorealistic samples.** (a) example of the airplane setting; (b) Example for the flower setting; (c) grid heatmap on `Qwen3.6-35B-A3B` (Qwen Team, 2026c). Heatmaps show the positive (red) and negative deviation (blue) for $\pm 3 std$ from the average performance over all patches. The number below the feature name indicates the spatial $mean \pm std$ accuracy.

B Comparison of Spatial Variance Patterns under Long(er) Context

In most experiments, we have only compared performance on three positions, as an exhaustive grid search would have been computationally infeasible. In this section, we want to estimate how the pattern of anisotropy changes when the resolution (sequence length) increases. Since this experiment would be infeasible at the rate of increase tested in §4.4.2, we at least want to estimate the trend by a comparison under $4\times$ of the image sequence length for `Qwen3.6-35B-A3B` (Qwen Team, 2026c) and `InternVL3.5-30B-A3B` (Wang et al., 2025b). Unlike previous experiments, we compute this result with batching, which introduces a slight noise into the results due to precision issues and sequence padding.

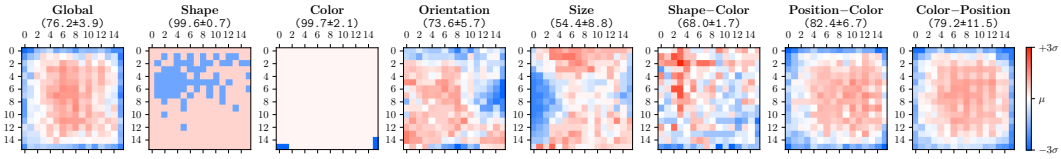
Note that these models use different strategies to tokenize high-resolution images. `InternVL` rescales and partitions inputs into tiles matching the encoder's native resolution of 448^2px. These tiles are encoded independently alongside a downsampled global thumbnail (Chen et al., 2024b; Wang et al., 2025b) using learnable position embeddings. In contrast, `Qwen` processes inputs at their native spatial resolution (Qwen Team, 2026a;b;c) using RoPE position embeddings (Heo et al., 2024).

The `Qwen` results in Figs. 10a and 10b show a highly similar pattern between both sequence lengths for all features, suggesting that the anisotropy pattern simply "scales" with the resolution (with minor fluctuations). We hypothesize that this is related to its usage of RoPE embeddings.

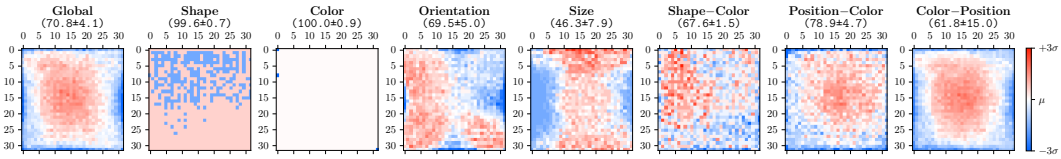
The spatial patterns on `InternVL` in Figs. 10c and 10d follow mixed trends: Some patterns are fairly similarly repeated for each tile (*e.g.*, shape, shape-color); others seem to "scale" similarly as on `Qwen` (*e.g.*, color-position); some even fully change (*e.g.*, color). Interestingly, some patterns (*e.g.*, orientation) may also show different patterns per tile, even if the overall pattern does not appear "scale". Some spatial biases almost fully vanish, as accuracy drops to random chance (*e.g.*, size).

C Language Priors

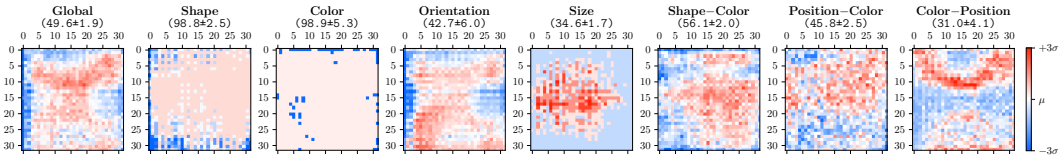
Our experiments focus on visual recognition while discarding the influence of language priors. In this section, we want to study how language priors may affect our results through three orthogonal axes: the order of answer options, the formulation of the prompt, and finally the parsing of responses.



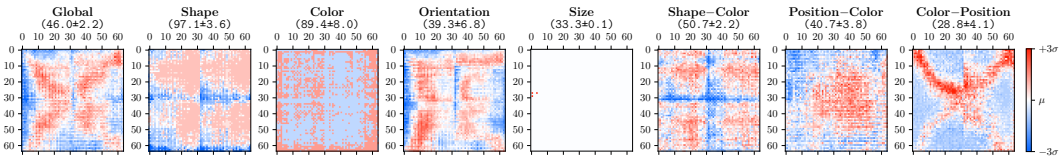
(a) `Qwen3.6-35B-A3B` (Qwen Team, 2026c): 256^2px (L=16^2)



(b) `Qwen3.6-35B-A3B` (Qwen Team, 2026c): 512^2px (L=32^2)



(c) `InternVL3.5-30B-A3B` (Wang et al., 2025b): 448^2px (L=32^2)



(d) `InternVL3.5-30B-A3B` (Wang et al., 2025b): 896^2px (L=64^2)

Figure 10: **Spatial distribution of feature fidelity with higher resolution.** We compute the accuracy for each feature for all positions in the patch grid for the minimum image sequence length and its quadruple on (a)+(b) `Qwen3.6-35B-A3B` (Qwen Team, 2026c) and (c)+(d) `InternVL3.5-30B-A3B` (Wang et al., 2025b). Heatmaps show the positive (red) and negative deviation (blue) for $\pm 3 std$ from the average performance over all patches. The number below the feature name indicates the spatial $mean \pm std$ accuracy.

C.1 Option Order Sensitivity

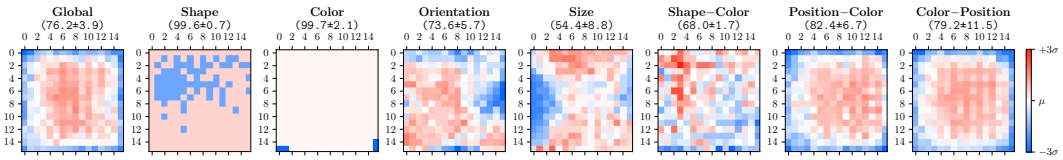
We first test the stability of our results with 3 additional runs of our on-grid experiments with `Qwen3.6-35B-A3B`, each with randomly shuffled options. Unlike the main results, these runs used batched inference, so differences from baseline cannot be attributed exclusively to option order.

Shuffling answer options preserves the baseline's spatial biases (Fig. 11): the central advantage and boundary disadvantage remain consistent across all three seeds; spatial patterns remain similar to the baseline. Global accuracy maps correlate strongly with the baseline, despite lower average accuracy. Orientation and size show greater changes in their detailed spatial patterns. These results overall support an invariance to option ordering.

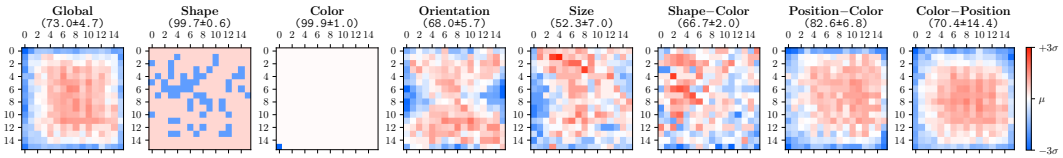
C.2 Prompt Sensitivity

To test anisotropy as a function of the prompt, we additionally conduct two runs of our on-grid experiments on `Qwen3.6-35B-A3B`, each with a manually verified rephrased but semantically similar prompt (generated by GPT 5.6 Sol Extra High).

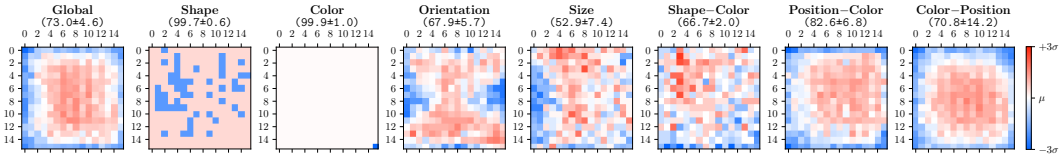
Compared to the option ordering experiment, which resulted in fairly similar anisotropy patterns, we find that even semantically similar prompts can induce different anisotropy patterns (Fig. 12). This does not invalidate our findings, but should serve as a caution that the observed anisotropy is partly



(a) Baseline



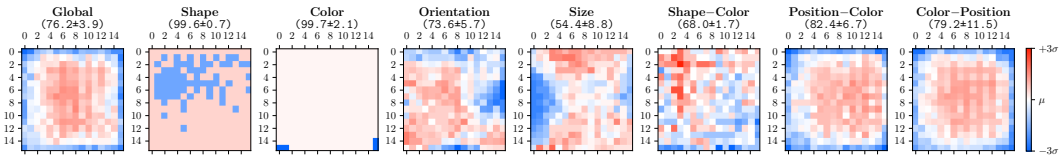
(b) Shuffled options (Seed 0)



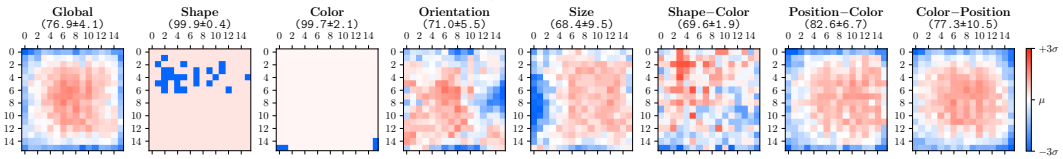
(c) Shuffled options (Seed 1)

Figure 11: **Anisotropy as an effect of shuffled options in the prompt.** (a) Anisotropy patterns of `Qwen3.6-35B-A3B` (Qwen Team, 2026c) as evaluated in the main experiments and (b)+(c) with shuffled options. Results were obtained using batched inference, causing noisy approximations of actual performance.

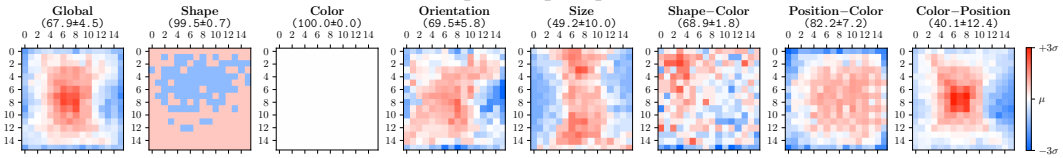
conditioned on the linguistic framing of the task and that evaluations should therefore account for prompt sensitivity and not mix different prompts.



(a) Baseline



(b) Rephrased prompts (1)



(c) Rephrased prompts (2)

Figure 12: **Anisotropy as an effect of rephrased prompts.** (a) Anisotropy patterns of `Qwen3.6-35B-A3B` (Qwen Team, 2026c) as evaluated in the main experiments and (b)+(c) with rephrased prompts. Results were obtained using batched inference, causing noisy approximations of actual performance.

C.3 Response Parsing

In our experiments, we parse the model responses based on simple heuristics. In Tab. 2, we verify the success rate of our heuristics for the generated responses for the experiments in §4.1 and find that the response rate is close to near-perfect rates. Therefore, it is suitable for our analysis. We do not claim generalization of our heuristics to other models or significantly different tasks or prompts.

Table 2: **Response parsing rate.** We report the fraction of parsed responses for each of our tested models on the VQA answers to our main experiments in §4.1.

	Rate of Parsed Responses [%]						
Model	Shape	Color	Orientation	Size	Shape-Color	Position-Color	Color-Position
`LLaVA1.5-7B`	99.99	100.00	98.17	100.00	100.00	100.00	100.00
`InternVL3.5-30B-A3B`	99.70	100.00	100.00	100.00	100.00	99.99	100.00
`Qwen3.6-35B-A3B`	100.00	99.87	100.00	100.00	100.00	99.99	100.00

D Reproducibility Statement

We utilized a stack consisting of Python 3.11.14 (van Rossum, 1995), `torch` 2.10.0 (Paszke et al., 2019), `transformers` 5.6.1 (Wolf et al., 2020), and CUDA 13.0. Experiments were performed on an internal cluster with single-node, multi-GPU NVIDIA L40S 48 GB, A100 PCIe 40 GB, or H100 HBM3 80 GB systems. The experiments do not require significant storage space, except to store the model weights. We utilized a system memory of 200-500 GB for each run. Some model files were hot-patched to support the utilized `transformers` version. Tab. 3 shows the consumed resources, excluding failed or non-utilized experiments.

To ensure full transparency and facilitate the reproduction of our results, we commit to releasing the complete codebase for our anisotropic patch evaluation upon paper acceptance.

Please refer to Tab. 4 to find the model slug and license for all models that we utilized in our experiments.

Table 3: **Cumulative overview of utilized resources per experiment and accelerator.**

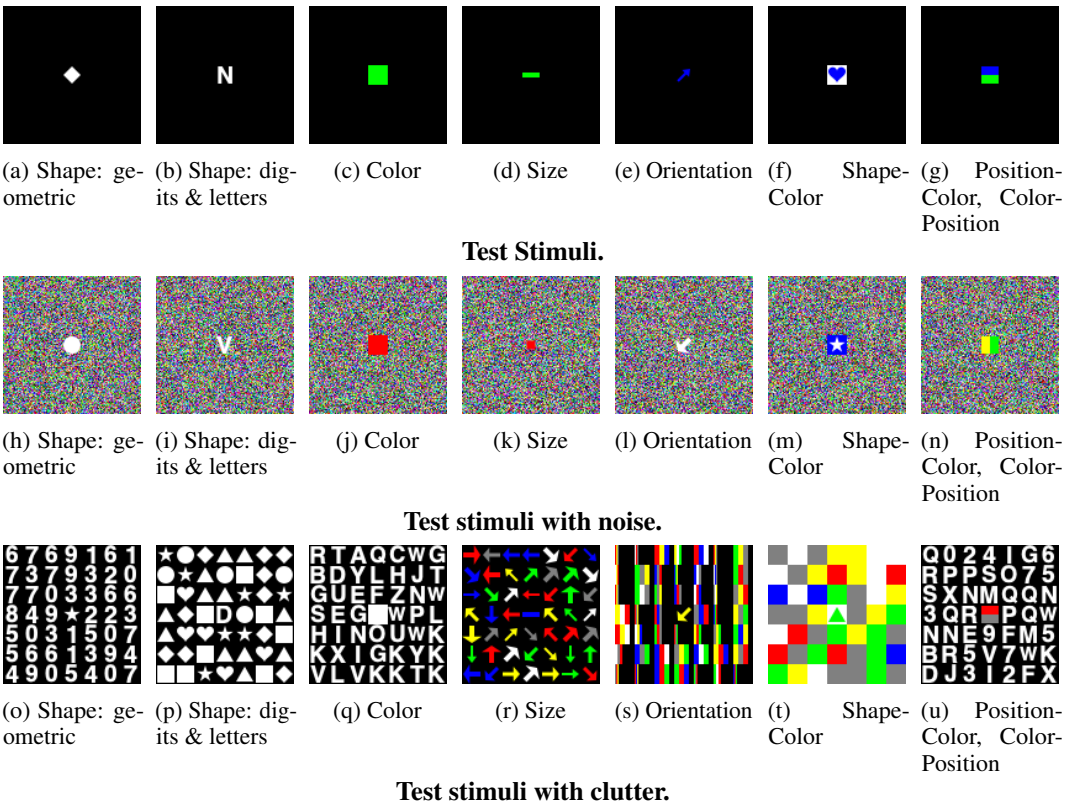
Experiment	**Accelerator**	**GPU hours**
Baseline (§4.1)	A100	2.2
On-grid (§4.1)	L40S	5.0
	H100	70.2
Fragmentation (§4.2)	L40S	3.1
	A100	2.1
	H100	3.1
Scale (§4.3)	L40S	3.8
	A100	2.7
	H100	0.5
Distractors (§4.4.1)	L40S	0.5
	A100	6.1
	H100	2.1
Sequence length (§4.4.2)	H100	18.6
Photorealistic tasks (§A)	A100	0.5
On-grid longer sequence length [batched] (§B)	H100	59.3
Total	L40S	12.4
	A100	13.6
	H100	153.8
	All GPUs	179.8

Table 4: **Model details.**

Model	HuggingFace Slug	License
LLaVA1.5-7B	llava-hf/llava-1.5-7b-hf	Llama 2 Community
InternVL3.5-30B-A3B	OpenGVLab/InternVL3_5-30B-A3B-HF	Apache 2.0
Qwen3.6-35B-A3B	Qwen/Qwen3.6-35B-A3B	Apache 2.0

E VISUALIZATION OF TEST STIMULI

We provide representative examples of our test stimuli in Fig. 13. The top row illustrates our baseline cases, where the feature patch is centered for simplicity; during evaluation, this placement is varied according to the specific experimental configuration. The middle row depicts stimuli containing uniform color noise as a distractor. Finally, the bottom row showcases stimuli with background clutter, in which all non-target patches are populated with unrelated features. In all cases, only a single patch contains the task-relevant features. Furthermore, Tab. 6 details the distractor combinations utilized in our clutter experiments. We ensure that distractors are related but disjoint from the target features specified in the prompt to prevent semantic overlap or task ambiguity.



(a) Shape: geometric (b) Shape: digits & letters (c) Color (d) Size (e) Orientation (f) Shape-Color (g) Position-Color, Color-Position

(h) Shape: geometric (i) Shape: digits & letters (j) Color (k) Size (l) Orientation (m) Shape-Color (n) Position-Color, Color-Position

(o) Shape: geometric (p) Shape: digits & letters (q) Color (r) Size (s) Orientation (t) Shape-Color (u) Position-Color, Color-Position

Figure 13: **Examples of test stimuli per feature.** For visualization purposes, the active patch is always placed in the center of the grid.

F LIST OF PROMPTS PER FEATURE

In the following, in Tab. 5 we provide the prompts used for evaluating the anisotropic feature predictions of our tested models.

Table 5: **Prompts for each Feature.** Each feature may have one or more prompts. Please note that we further divide the features into subcategories (*e.g.*, shape consists of geometric shapes, letters, and digits, each with its own prompt). Values separated by | denote options. Values between %% are inferred from the ground truth. Prompts in gray are prompts for the additional experiments in the appendix.

Task	Prompts/ Image	Prompts
Shape	1	**For Shapes/Outlines:** What shape is in this image? Options: Circle, Square, Triangle, Heart, Star, Diamond. Give a single word answer. **For Digits:** What digit is in this image? Options: 0-9. Give a single digit answer. **For Letters:** What letter is in this image? Options: A-Z. Give a single letter answer.
Color	1	What is the color of the small patch? Options: Red, Green, Blue, Yellow, White, Gray. Give a single word answer.
Size	1	What is the side ratio of the rectangle in this image? Options: Wide, Tall, Square. Answer with exactly one word.
Orientation	1	What direction is the arrow pointing to? Options: North, South, East, West, Northeast, Northwest, Southeast, Southwest. Give a single word answer.
Shape-Color	1	What color is the {%shape_type%} in the image? Options: Red, Green, Blue, Yellow, White, Gray. Give a single word answer.
Position-Color	2	What color is the {left | right} half of the square in this image? Options: Red, Green, Blue, Yellow, White, Gray. Give a single word answer. What color is the {top | bottom} half of the square in this image? Options: Red, Green, Blue, Yellow, White, Gray. Give a single word answer.
Color-Position	2	Which half of the square is {%left_color% | %right_color%}? Options: Top, Bottom, Left, Right. Give a single word answer. Which half of the square is {%top_color% | %bottom_color% }? Options: Top, Bottom, Left, Right. Give a single word answer.
Plane: Color	1	What color is the plane in the image? Options: Red, Green, Blue, Yellow, White, Black. Give a single word answer.
Plane: Direction	1	What direction is the plane flying to in the image? Options: Up, Down, Left, Right. Give a single word answer.
Flower: Petal-Color	1	What color are the petals of the flower in the image? Options: Red, Green, Blue, Yellow, White, Black. Give a single word answer.
Flower: Center-Color	1	What color is the center of the flower in the image? Options: Red, Green, Blue, Yellow, White, Black. Give a single word answer.

G Full Evaluation Results

In the following, we provide the full (non-aggregated) evaluation plots for our experiments in §4. Figs. 14 to 17 all follow the same layout, with plots organized in a model (rows) × features (columns) grid, where we measure the performance of the respective visual factor on three positions (X-axis). The first row shows results on `Qwen3.6`, the second for `InternVL3.5`, and the third for `LLaVA1.5`, besides Fig. 17, where we did not include `LLaVA1.5` as it does not support higher

Table 6: **Distractors per Feature.**

Feature	Distractor
Shape: Geometric	Shape: Digits
Shape: Digit	Shape: Geometric
Shape: Letter	Shape: Geometric
Color	Letters
Orientation	Vertical color bars of different width
Size	Orientation
Shape-Color	Color
Position-Color & Color-Position	Shape: Letter + digits

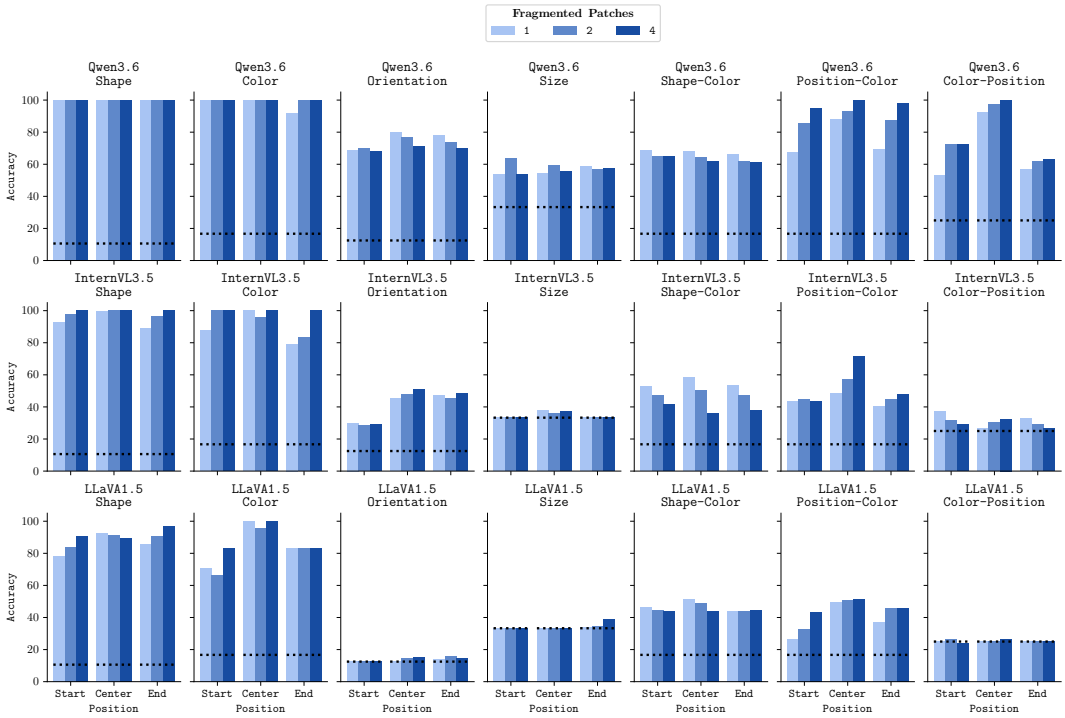


Figure 14: **Feature recognition under fragmentation (not aggregated).** Dotted horizontal lines show the random chance.

resolutions. Dotted lines indicate the random chance. Fig. 14 depicts the detailed evaluation on the fragmentation experiments described in §4.2, with fragmentation into 1, 2 or 4 patches. Fig. 15 depicts the detailed evaluation on the feature recognition under scaling as described in §4.3 using scales $s \in \{1, 2, 4, 8\}$. Fig. 16 depicts the detailed evaluation on the distractor experiments described in §4.4.1, with no distractors marked as "base", "noise" distractor, and "clutter". Fig. 17 depicts the detailed evaluation on the long context experiments described in §4.4.2, baseline sequence length marked with "base" and increased sequence length marked with "long".

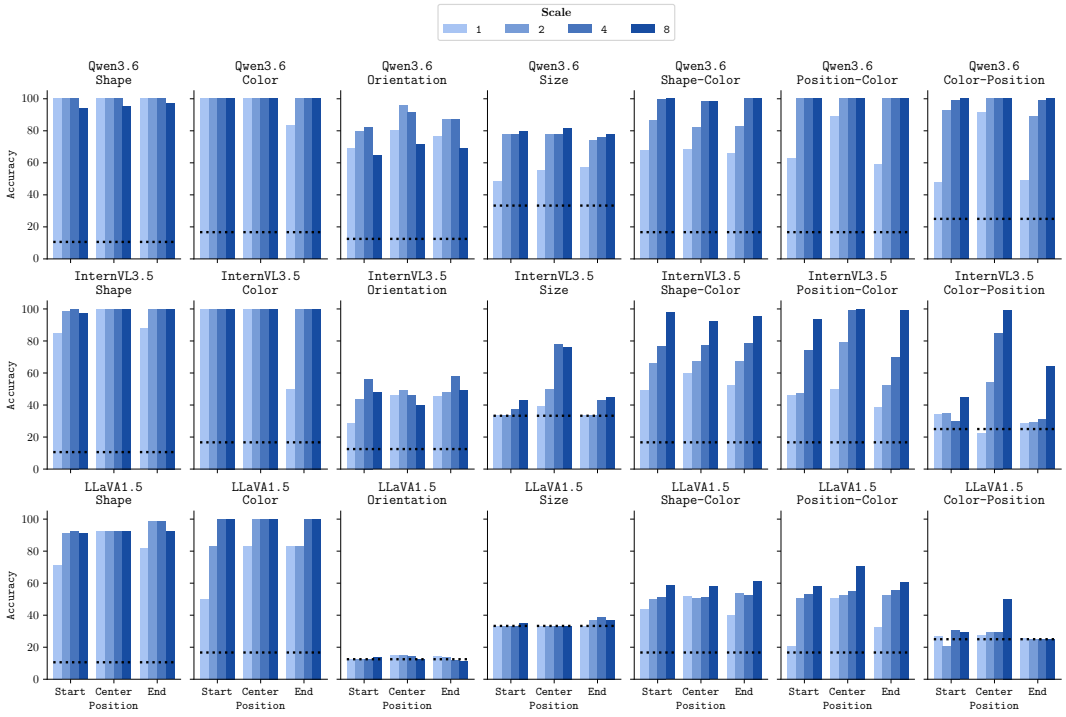


Figure 15: **Feature recognition under scale (not aggregated).** Dotted horizontal lines show the random chance.

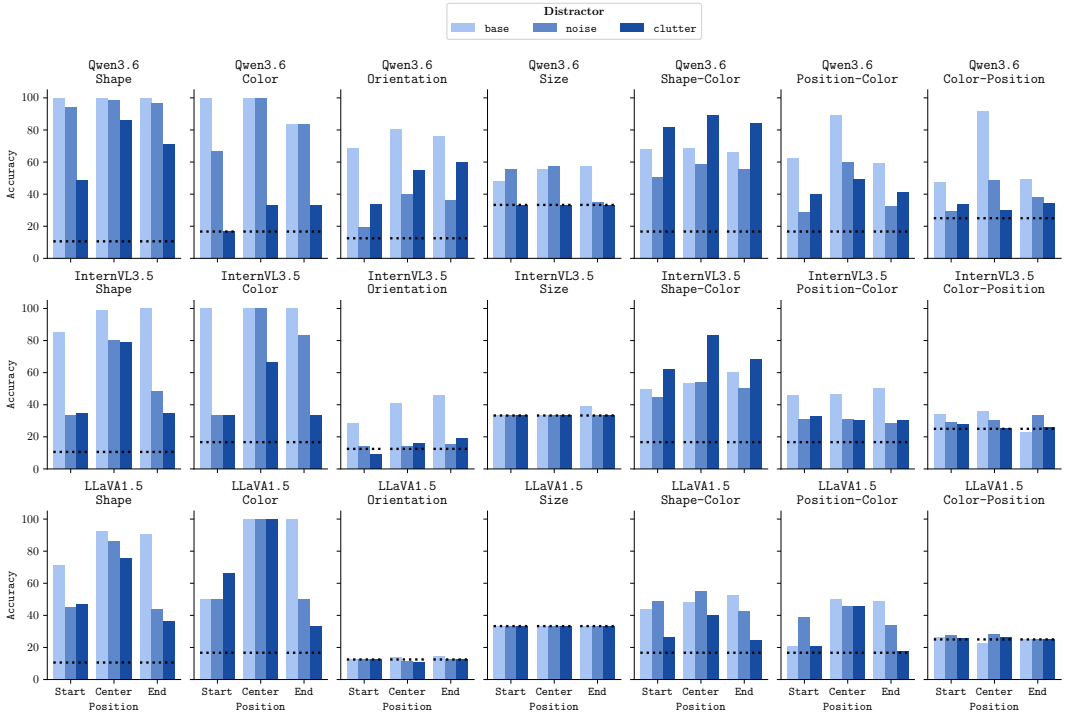


Figure 16: **Feature recognition under visual distraction (not aggregated).** Dotted horizontal lines show the random chance.

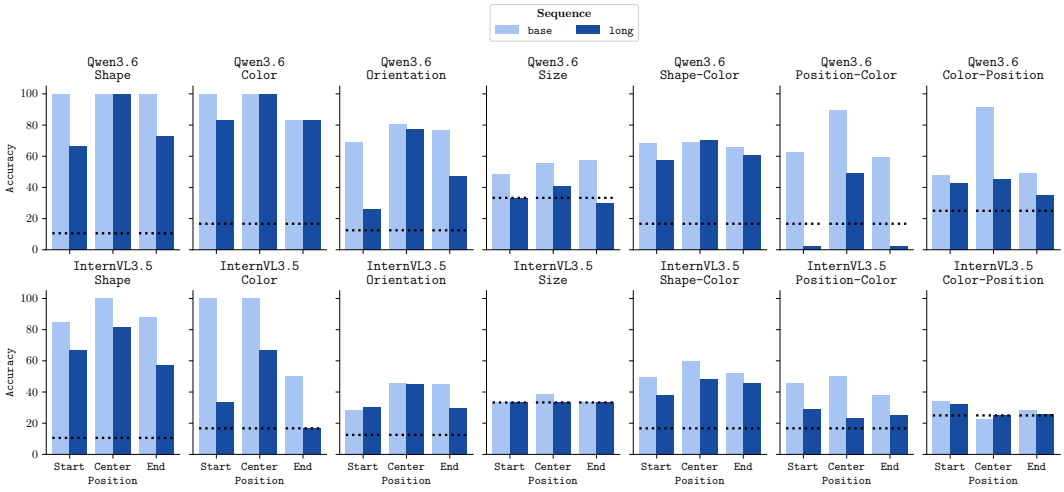


Figure 17: **Feature recognition under long context (not aggregated).** Dotted horizontal lines show the random chance.