FREQUENCY BOUNDARIES OF VISUAL SEMANTICS FOR GROUNDED ANSWERING IN FROZEN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models answer questions across diverse visual tasks, yet how image frequency bands affect answer correctness and semantic content remains unclear. We introduce a paired analysis framework based on independent band suppression to characterize the frequency boundaries of visual semantics in frozen models. Holding the model and question fixed, we independently generate eight radial band-stop variants of each image and compare original and filtered responses using five semantic labels based on the question and reference answers, distinguishing preservation, degradation, and error correction. Across four tasks and five models from two families, degradation follows task-dependent frequency-band patterns shared across models. Some initially incorrect answers also become correct under band suppression. For shared instances answered correctly by all five models on unfiltered images, InternVL3 has significantly lower correct-to-incorrect transition rates than Qwen2.5-VL across multiple bands in text-centric tasks. Within families, larger models show similar reductions, mainly in document understanding. Human validation yields 91.72% weighted response-level agreement between final automatic labels and human judgments. These findings link suppressed bands to the direction and type of semantic transition, revealing how frequency boundaries vary with task, family, and scale.

1 INTRODUCTION

Vision-language models (VLMs) combine visual representations with language generation to answer questions about images (Li et al., 2023; Liu et al., 2023a). Their applications span natural scenes, scene text, documents, and medical images (Antol et al., 2015; Singh et al., 2019; Mathew et al., 2021; Liu et al., 2021). Correct answers on unfiltered images, however, do not reveal their sensitivity to changes in visual information. Generated content can conflict with visual evidence (Leng et al., 2024; Hwang et al., 2026). It is unclear whether answers remain correct, specific, and complete when selected spatial frequency components are suppressed.

Spatial frequency provides a controllable dimension of visual input. Image recognition studies show that even correct predictions can rely on cues different from those used by humans (Geirhos et al., 2018). Frequency analyses further show that model sensitivity varies across frequency components and that retaining or selecting different components affects recognition performance (Yin et al., 2019; Xu et al., 2020). These findings do not establish how frequency changes affect answer quality in open-ended visual question answering. The effects of band suppression may also depend on the information each task requires. Direct evaluation is therefore needed to determine whether an answer remains correct, loses information, or diverges from its reference when a particular band is suppressed.

Frequency information has also been used to study VLM generation. Contrastive decoding and perturbation methods use model outputs or visual features under multiple frequency conditions to reduce object hallucinations (Liu et al., 2025; Li et al., 2025). These studies link frequency manipulation to generation errors but primarily aim to improve responses rather than track answer quality across frequency conditions for the same image and question. Lower aggregate hallucination rates do not reveal whether individual answers still agree with their references, lose information,

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.

or introduce conflicting content. Paired comparisons of original and filtered responses are therefore needed, but binary correctness is insufficient. An answer may remain compatible with a reference while becoming less specific, or correctly answer some requested components while omitting others. FaithScore evaluates visual faithfulness through atomic facts, whereas VALOR-EVAL assesses faithfulness and information coverage separately, showing why a single score cannot capture all aspects of answer quality (Jing et al., 2024; Qiu et al., 2024). Aggregate scores can also conceal opposing changes under frequency interventions: correct answers may become incorrect, while incorrect answers may become correct. Evaluation must therefore assess each answer against its reference and track the direction and type of transition.

To address this problem, we compare frozen VLM responses to the same question given an original image and eight independently generated radial band-stop variants. We define five semantic labels based on the question and dataset reference answers and assign one to each original and filtered response. The resulting label transitions characterize the frequency boundaries of visual semantics, distinguishing preservation, degradation, and error correction without assuming that the original response is correct.

We evaluate five model configurations from Qwen2.5-VL and InternVL3 (Bai et al., 2025; Zhu et al., 2025) on GQA, TextVQA, DocVQA, and SLaKE (Hudson & Manning, 2019; Singh et al., 2019; Mathew et al., 2021; Liu et al., 2021). Our contributions are threefold. First, our paired framework links independent band suppression to response changes with the model and question fixed, characterizing frequency boundaries in frozen VLMs. Second, five semantic states based on reference answers distinguish preservation, degradation, and correction through directed transitions between original and filtered response labels. We assess annotation agreement with human judgments. Third, our analysis identifies task-dependent degradation patterns across bands that are shared across models, alongside transition differences associated with family and scale. Together, these results provide evidence across tasks and models for the frequency boundaries of visual semantics.

2 Related Work

Image Frequency and Vision-Language Models. Frequency modulation has been used to improve understanding of nonnatural images and adapt VLM visual representations using question text (Lin et al., 2025; Ouyang et al., 2026). Other work targets object hallucinations in generated responses. Visual contrastive decoding adjusts generation using output distributions from original and distorted images (Leng et al., 2024). Contrastive decoding across multiple frequencies uses image conditions with high and low frequency content (Liu et al., 2025), while frequency perturbation methods perturb high and low frequency features during adversarial training (Li et al., 2025). These approaches aim to improve representations or reduce object hallucinations rather than suppress individual bands to examine their effects on VLM answers. We instead generate eight independent radial band-stop variants of the same image, keeping the model and question fixed to identify which bands alter responses when suppressed.

Semantic Quality of Visual Answers. Visual question answering models generate responses from images and questions, with correctness evaluated against dataset reference answers (Antol et al., 2015; Hudson & Manning, 2019). Open-ended visual answers, however, have semantic properties beyond correctness. FaithScore decomposes responses into atomic facts and verifies their faithfulness against the image (Jing et al., 2024), whereas VALOR-EVAL evaluates visual faithfulness and information coverage separately (Qiu et al., 2024). These approaches assess reference agreement, visual faithfulness, and information coverage for individual answers, rather than paired changes for the same question under altered frequency content. Such comparisons require evaluating original and filtered responses under consistent semantic criteria.

Response Evaluation with LLMs. LLM-as-a-Judge methods evaluate generated responses using scoring rubrics, reference answers, or pairwise comparisons (Liu et al., 2023b; Zheng et al., 2023; Kim et al., 2024). Agreement with human judgments varies by task and evaluation criterion (Bavaresco et al., 2025). We independently label original and filtered responses using a fixed rubric of five semantic classes, the same question, and the same references, and check agreement with human annotations on a sampled set. Pairing the labels then links semantic transitions to individual band suppression, characterizing the frequency boundaries of visual semantics.

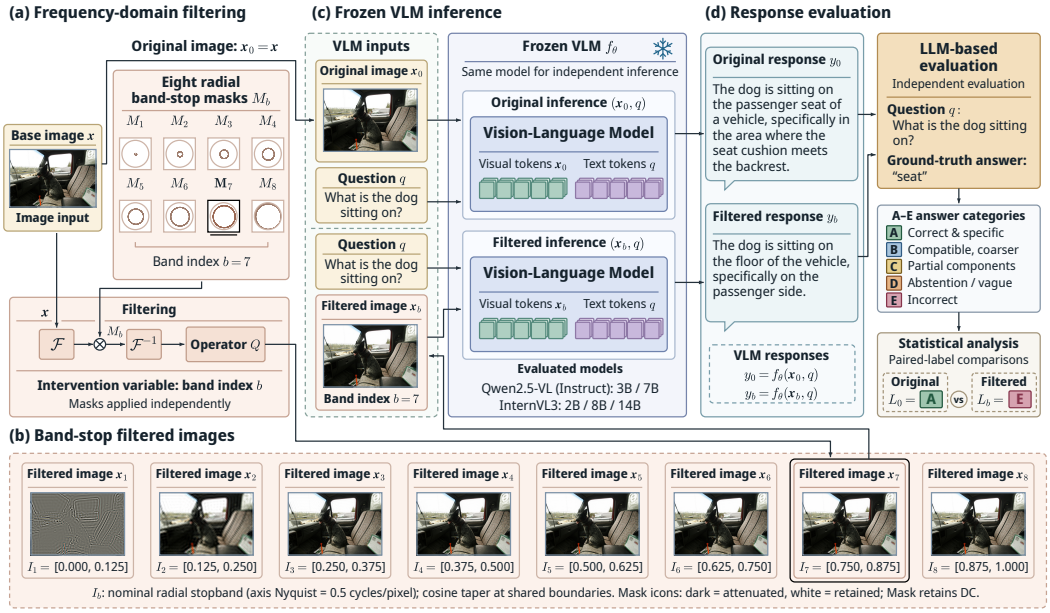


Figure 1: Paired analysis of visual answers under individual band suppression. Starting from the same original image, (a) eight radial band-stop masks define the filtering conditions, and (b) applying each mask independently produces eight images, each with a different band suppressed. For each band, (c) the original image and its filtered counterpart are processed separately by the same frozen VLM with the same question. Finally, (d) an evaluation model uses the question and reference answers to assign each response a semantic label independently. Each of the eight band labels is then paired with the original label to compare semantic states. The example illustrates the seventh band.

3 METHOD

To characterize frequency boundaries in frozen VLMs, we suppress individual bands to create comparable image conditions and evaluate response semantics against dataset reference answers. Figure 1 presents the complete pipeline. We independently generate eight filtered images from the same original and obtain responses to all nine images with the question and VLM fixed. Evaluating each response using the same question and references yields one original label and eight band labels. Pairing each band label with the original label gives eight comparisons of semantic preservation and change under individual band suppression.

3.1 INDEPENDENT BAND SUPPRESSION

To analyze response changes under band suppression, we independently generate eight filtered versions of the same base image. Appendix A examines the relationship between band count and spectral binning error. Let $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ denote the base RGB image with height H and width W, and $\mathbf{x}_0 = \mathbf{x}$ its unfiltered form. For each $b \in \{1, \ldots, 8\}$, applying the bth radial band-stop mask to the spectrum of $\mathbf{x}_0$ yields the filtered image $\mathbf{x}_b$.

Let $\mathcal{F}$ denote the two-dimensional discrete Fourier transform applied to each color channel, with f_x and f_y denoting horizontal and vertical frequency coordinates. Normalizing each axis by its Nyquist frequency $f_N = 0.5$ cycles/pixel gives the radial coordinate $\rho(f_x, f_y) = \sqrt{(f_x/f_N)^2 + (f_y/f_N)^2}$. Within $0 \leq \rho \leq 1$, the nominal bandwidth is $\Delta = 1/8$, and the lower and upper bounds are $\ell_b = (b-1)\Delta$ and $u_b = b\Delta$. The nominal radial stopband is thus $I_b = [\ell_b, u_b]$. A fixed transition half-width $\tau = 0.1\Delta$ provides smooth attenuation around shared band boundaries. For a signed radial offset $z \in \mathbb{R}$ from a boundary, let $\mathrm{clip}(z, -\tau, \tau) = \min\{\tau, \max\{-\tau, z\}\}$. The resulting raised-cosine step function is

$$s_\tau(z) = \tfrac{1}{2}\left[1 + \sin\left(\frac{\pi\ \mathrm{clip}(z, -\tau, \tau)}{2\tau}\right)\right]. \tag{1}$$

For $0 < \rho \leq 1$, the suppression weight W_b and band-stop mask M_b are defined as

$$W_b(\rho) = \begin{cases} 1 - s_\tau(\rho - u_b), & b = 1, \\ s_\tau(\rho - \ell_b) - s_\tau(\rho - u_b), & 2 \leq b \leq 7, \\ s_\tau(\rho - \ell_b), & b = 8, \end{cases} \qquad M_b(\rho) = 1 - W_b(\rho). \tag{2}$$

Setting $W_b(\rho) = 0$ for $\rho = 0$ or $\rho > 1$ preserves the DC coefficient and coefficients outside the unit disk at the mask stage, before clipping and rounding. Elsewhere, $M_b = 0$ removes the corresponding Fourier coefficient, $M_b = 1$ retains it, and $0 < M_b < 1$ provides cosine attenuation near shared band boundaries. Evaluating $M_b(\rho(f_x, f_y))$ at each grid point produces a mask aligned with the image spectrum, also denoted by M_b. Applying this mask to all three color channels and reconstructing the image gives

$$\mathbf{x}_b = Q\big(\mathcal{F}^{-1}[M_b \odot \mathcal{F}(\mathbf{x}_0)]\big), \qquad b \in \{1, \ldots, 8\}. \tag{3}$$

Here, $\mathcal{F}^{-1}$ denotes the inverse Fourier transform applied to each channel and $\odot$ denotes pointwise multiplication over frequency locations. The reconstruction operator Q takes the real part, clips it to the valid pixel range, and rounds to the nearest integer to produce an RGB image with eight bits per channel. All conditions use the same nominal bandwidth Δ, transition half-width τ, mask depth 1, and reconstruction rule Q. Only the suppressed band index b varies. Each $\mathbf{x}_b$ is generated independently from $\mathbf{x}_0$, giving eight image pairs $(\mathbf{x}_0, \mathbf{x}_b)$.

3.2 Paired Inference with Frozen Models

For each question instance, let q denote the dataset question and f_θ a VLM with parameters θ, evaluated with fixed task prompts and decoding settings. We provide the original image $\mathbf{x}_0$ and each filtered image $\mathbf{x}_b$ separately to the model with the same question q, yielding

$$y_0 = f_\theta(\mathbf{x}_0, q), \qquad y_b = f_\theta(\mathbf{x}_b, q), \quad b \in \{1, \ldots, 8\}. \tag{4}$$

Model weights, task prompt formats, and deterministic decoding settings remain fixed across conditions. Each condition is evaluated independently; the model receives no images or responses from other conditions. Each instance yields one original and eight filtered responses, forming eight pairs (y_0, y_b) with y_0 as their common baseline.

3.3 Semantic Annotation Using Reference Answers

To compare response semantics across original and filtered conditions, we annotate y_0 and each y_b separately using question q and its reference answer set $\mathcal{R}_q$. Multiple references are treated as alternatives, and matching one suffices. For each response y_i, $i \in \{0, \ldots, 8\}$, annotation run $s \in \{1, 2, 3\}$ of model J yields

$$\widetilde{L}_{i,s} = J_s(q, \mathcal{R}_q, y_i) \in \mathcal{Y} \cup \{\perp\}, \qquad \mathcal{Y} = \{A, B, C, D, E\}. \tag{5}$$

Here, $\perp$ denotes the absence of a valid label. Model J receives only the original dataset question, its reference answers, and one response, and assesses response content against shared references. Final labels L_i are obtained by majority voting in Section 4.2, with valid states defined in Table 1. Appendix B gives the complete prompt.

Table 1: Semantic answer states assigned using reference answers. The definitions follow the prompt in Appendix B.

Label	Semantic state	Definition
A	Correct and specific	Correct, complete, and sufficiently specific.
B	Compatible but coarser	Compatible with a reference answer, but coarser on the same requested attribute.
C	Incomplete multipart answer	Some separately requested components are correct, while others are omitted.
D	Abstention or nonanswer	An abstention or a vague response that does not answer the question.
E	Incorrect or unrelated	Wrong or conflicting requested content, an unrelated answer, or an empty response.

3.4 Semantic Transitions and Frequency Boundaries

Semantic labels assigned using the same question and reference answers provide a consistent basis for comparing original and filtered responses. We compare these labels to determine whether responses retain the requested information after band suppression. For resolved labels $L_0, L_b \in \mathcal{Y}$, the directed pair for band b is

$$T_b = (L_0, L_b) \in \mathcal{Y} \times \mathcal{Y}. \tag{6}$$

This pairing preserves the direction of change, linking each suppressed band to losses of specificity or information, abstention, or errors in initially correct responses, as well as corrections of initially incorrect responses. Comparing these outcomes across bands identifies conditions associated with semantic preservation or particular transitions. These conditions and the direction and type of transition define the frequency boundaries of VLM answers.

4 Experiments

We evaluate response changes under band suppression on GQA, TextVQA, SLaKE, and DocVQA (Hudson & Manning, 2019; Singh et al., 2019; Liu et al., 2021; Mathew et al., 2021), covering compositional visual reasoning, scene text question answering, medical visual question answering, and document understanding, respectively. Five frozen models from Qwen2.5-VL and InternVL3 are evaluated (Bai et al., 2025; Zhu et al., 2025). For each instance, we obtain original and filtered responses with the question and inference settings fixed, annotate them using dataset reference answers, and consolidate labels by majority voting before comparing paired transitions. Appendix C describes the datasets.

4.1 Models and Inference Settings

The five evaluated VLMs are Qwen2.5-VL-3B, Qwen2.5-VL-7B, InternVL3-2B, InternVL3-8B, and InternVL3-14B (Bai et al., 2025; Zhu et al., 2025), spanning two families and multiple parameter scales. Models remain frozen, with question formats and generation settings fixed for each model and task. Decoding selects the most probable token at each step without sampling. Under our settings, the same model and input produce identical responses across seeds, so each image condition contributes one response. Images are converted to RGB. Those exceeding 262,144 pixels are proportionally downsampled to fit this limit using Lanczos interpolation; the rest retain their original dimensions. Frequency interventions use the eight radial band-stop masks defined in Section 3.1, with spectral analysis in Appendix A. Each filtered image is generated independently from the original and processed using the same question and decoding settings.

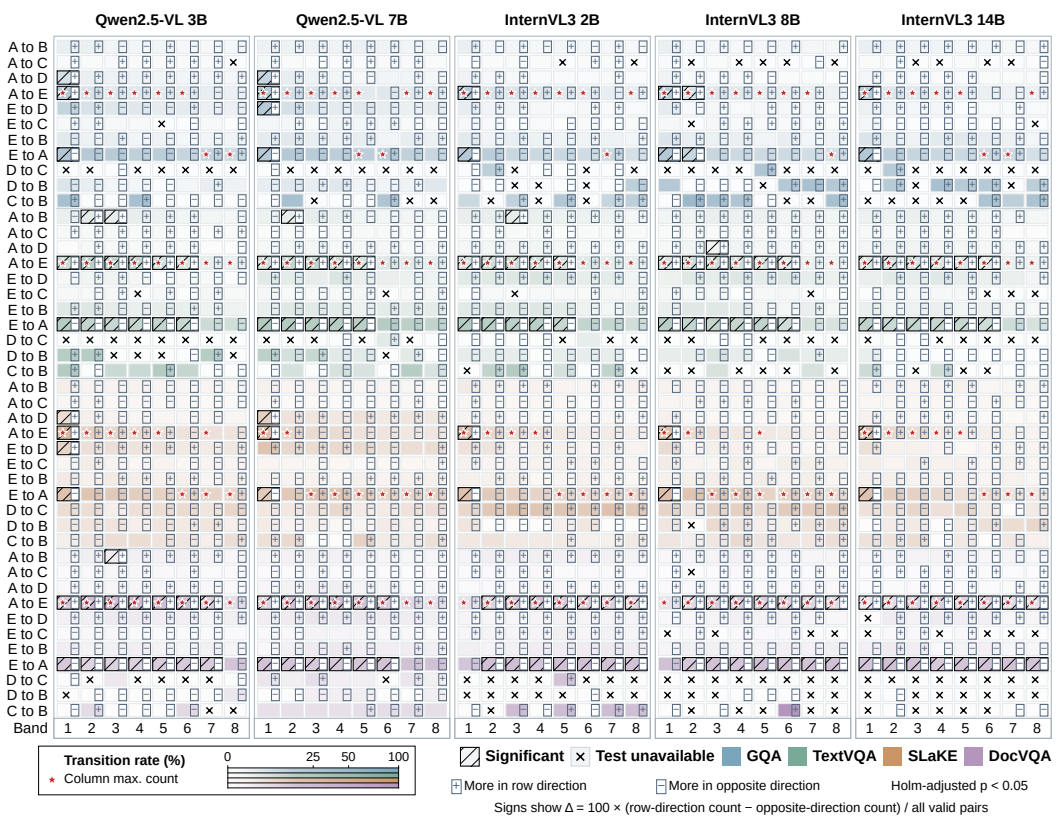


Figure 2: Semantic transitions under band suppression across tasks. Color intensity indicates the conditional transition rate. Red stars mark the largest observed count among the displayed directions within each task, model, and band, including ties. Black outlines and hatching indicate a significant difference between forward and reverse joint transition proportions after Holm correction ($p < 0.05$). Signs indicate the direction of the difference, and crosses denote unavailable tests. Appendix E.3 provides the definitions.

4.2 Semantic Annotation and Majority Voting

We use gpt-5.6-luna (OpenAI, 2026) to assign one of the five semantic labels defined in Section 3.3 using the question, reference answers, and one response, following the prompt in Appendix B. Within each model and dataset, labels $\widetilde{L}_{i,s}$ from three seeds are combined for each response. At least two matching valid labels determine $L_i \in \mathcal{Y}$. Otherwise, $L_i = \perp$ denotes an unresolved result, including cases with three different labels. Analysis includes only pairs with $L_0, L_b \in \mathcal{Y}$.

4.3 Evaluation Metrics and Comparison Design

Semantic transition analysis uses all valid pairs after majority voting and covers eleven directions: A to B, C, D, or E; E to D, C, B, or A; D to C or B; and C to B. For each direction, the denominator is the number of valid pairs with the designated initial label. The numerator counts those pairs in which the filtered response receives the target label and differs in text from the original response.

Analyses cover task and band, model family, and model scale comparisons across all eight bands. Task and band analyses include paired comparisons with the original condition, pairwise band comparisons within tasks, task comparisons at each band, and interactions between task and band. Family comparisons use shared instances with the designated original label and valid labels at all eight bands for all five models. We average model rates equally within each family before comparing InternVL3 with Qwen2.5-VL. Scale comparisons apply the same criteria within each family to compare larger and smaller models. For each model and dataset, overall results are equally weighted

means of the eight band rates, each using its defined eligible set. Effect sizes are rate differences for ordinary comparisons and differences in band contrasts between tasks for interactions, expressed in percentage points.

Effect sizes are computed from all eligible pairs. Pointwise 95% percentile confidence intervals are obtained from 10,000 bootstrap resamples of image clusters, preserving the pairing of questions, bands, and models within each image. Significance is assessed using cluster jackknife variance estimates and approximate two-sided t tests. Paired tests comparing forward and reverse joint transition proportions form a separate Holm correction family. The remaining comparisons are corrected separately within the task and band, model family, and model scale analyses, with adjusted $p < 0.05$ indicating significance. Appendix D provides detailed definitions of transition rates, comparison sets, and statistical inference.

4.4 Human Validation Design

Human validation uses twenty strata defined by five models and four datasets. We randomly sample fifty distinct images per stratum, selecting one question and one band condition per image with approximately balanced coverage of all eight bands. Original and filtered responses yield 1,000 pairs and 2,000 response records. Merging records with identical questions, reference answers, and response text leaves 1,402 unique annotation items. A further 21 responses from rare classes provide supplementary checks, giving 1,423 annotated items in total. Items are presented in random order. The annotator sees only the question, reference answers, and one response, and applies the five semantic criteria in Section 3.3. Human labels are then mapped back to the source response pairs and compared with the final automatic labels obtained by majority voting.

Using human labels as the reference, we report response-level and pair-level agreement, nominal Cohen's κ, and macro-F1 across the five classes. Pair-level agreement requires both response labels to match the human labels. Estimates use inverse sampling probability weights. Within each model and dataset stratum, 10,000 bootstrap resamples of pairs grouped by image yield 95% percentile confidence intervals. Appendix D.2 gives the eligibility criteria, weights, and metric definitions.

5 Results

5.1 Task Dependence of Band Transitions

Suppressing the first band can substantially alter an image's overall appearance. In Figure 1(b), the example is nearly grayscale and dominated by local contours, making the original colors and object shapes harder to recognize. Correct-to-incorrect transitions, however, are not confined to conditions with such pronounced visual changes. The paired tests in Figure 2 show that these transitions significantly outnumber reverse transitions across all five models in bands 1 to 5 on TextVQA and bands 2 to 6 on DocVQA. GQA and SLaKE show the same degradation tendency in band 1, although some initially incorrect answers also become correct. For some models, error correction is the most frequent observed transition in bands 6 to 8 on GQA and bands 3 to 8 on SLaKE. In these conditions, incorrect-to-correct transitions outnumber correct-to-incorrect transitions, but the paired differences are not significant. Pairwise band comparisons further characterize this task dependence. Across all five models, correct-to-incorrect transition rates are significantly higher in band 2 than in band 8 on GQA and SLaKE, in bands 2 to 4 than in bands 7 and 8 on TextVQA, and in bands 3 to 5 than in band 1 on DocVQA (Appendix Figure 5). Incorrect-to-correct transition rates are significantly higher in band 1 than in band 8 for all three InternVL3 models on GQA and for Qwen2.5-VL-3B, InternVL3-8B, and InternVL3-14B on SLaKE. These correction rates are also significantly higher in band 2 than in band 8 for InternVL3-2B on GQA and InternVL3-14B on SLaKE (Appendix Figure 5). Together, the distributions of degradation and correction across tasks and suppressed bands characterize the frequency boundaries of visual semantics (Appendix E.2).

5.2 Model Family Differences across Datasets

Under identical band suppression, correct-to-incorrect transition rates differ between model families by task (Figure 3). For shared instances answered correctly by all five models on unfiltered images, InternVL3 has significantly lower rates than Qwen2.5-VL in bands 1 to 7 on DocVQA and bands 1

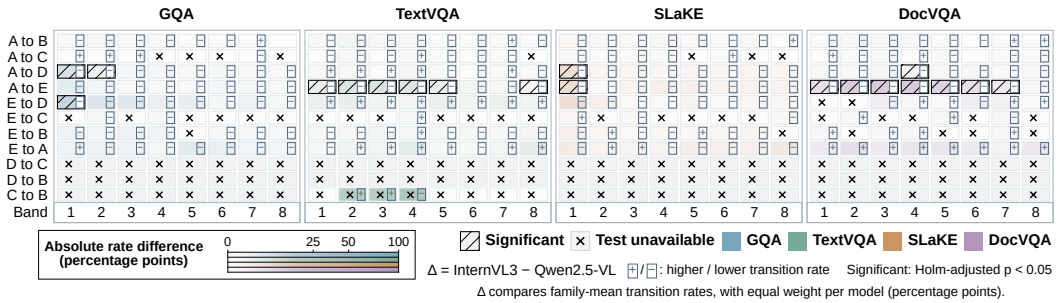


Figure 3: Model family differences on shared instances. Rates are averaged equally across models within each family before subtracting the Qwen2.5-VL rate from the InternVL3 rate. Color intensity indicates the absolute difference, and signs indicate its direction. Black outlines and hatching mark significant differences after Holm correction within the family analysis ($p < 0.05$). Crosses denote unavailable tests.

to 5 and 8 on TextVQA. On SLaKE, this difference is significant only in band 1. Averaged across the eight bands, the rate differences for InternVL3 relative to Qwen2.5-VL are -7.55 percentage points on DocVQA and -2.68 on TextVQA (95% CIs $[-8.48, -6.66]$ and $[-3.19, -2.16]$, respectively, both $p < 0.001$). These family differences in correct-to-incorrect transition rates vary by task and suppressed band, revealing distinct frequency boundaries.

5.3 Model Scale and Semantic Transitions

Within families, differences in correct-to-incorrect transition rates are concentrated on DocVQA (Figure 4). Qwen2.5-VL-7B has significantly lower rates than 3B in bands 1 to 7. For InternVL3, rates are significantly lower for 14B than for 2B in bands 2 to 8, whereas 8B has significantly lower rates than 2B only in band 4. Other tasks show more localized significant differences. On TextVQA, Qwen2.5-VL-7B has significantly lower rates than 3B in bands 2 and 4. Corresponding differences occur between InternVL3-14B and 2B in bands 4 and 6 on GQA, and between Qwen2.5-VL-7B and 3B in band 1 on SLaKE. Larger models thus show lower correct-to-incorrect transition rates across multiple bands on DocVQA, with significant differences elsewhere restricted to specific model pairs and bands.

5.4 Human Annotation Validation

In the primary human validation sample, final labels from majority voting yield weighted agreement with human judgments of 91.72% for individual responses and 87.73% for pairs, where both original and filtered response labels must match human judgments. These results support paired semantic transition analysis. Appendix D.2 reports all metrics, per-class results, and confidence intervals.

6 Discussion

Frequency boundaries of visual semantics vary by task. TextVQA and DocVQA are both text-centric tasks but show different degradation patterns across bands. These task-specific patterns are shared across all five models. On GQA and SLaKE, degradation and correction coexist, showing that suppressing the same band can affect initially correct and incorrect answers differently. Frequency boundaries therefore specify both the band conditions and semantic directions of response changes, directly linking image frequency information to answering behavior.

Model families differ in their correct-to-incorrect transition rates after band suppression. For shared instances answered correctly by all five models on unfiltered images, InternVL3 has significantly lower correct-to-incorrect transition rates than Qwen2.5-VL across multiple bands in scene text and document understanding. Thus, even when all models answer correctly on unfiltered images, the families have distinct frequency boundaries, reflected in different proportions of initially correct answers becoming incorrect under identical suppression.

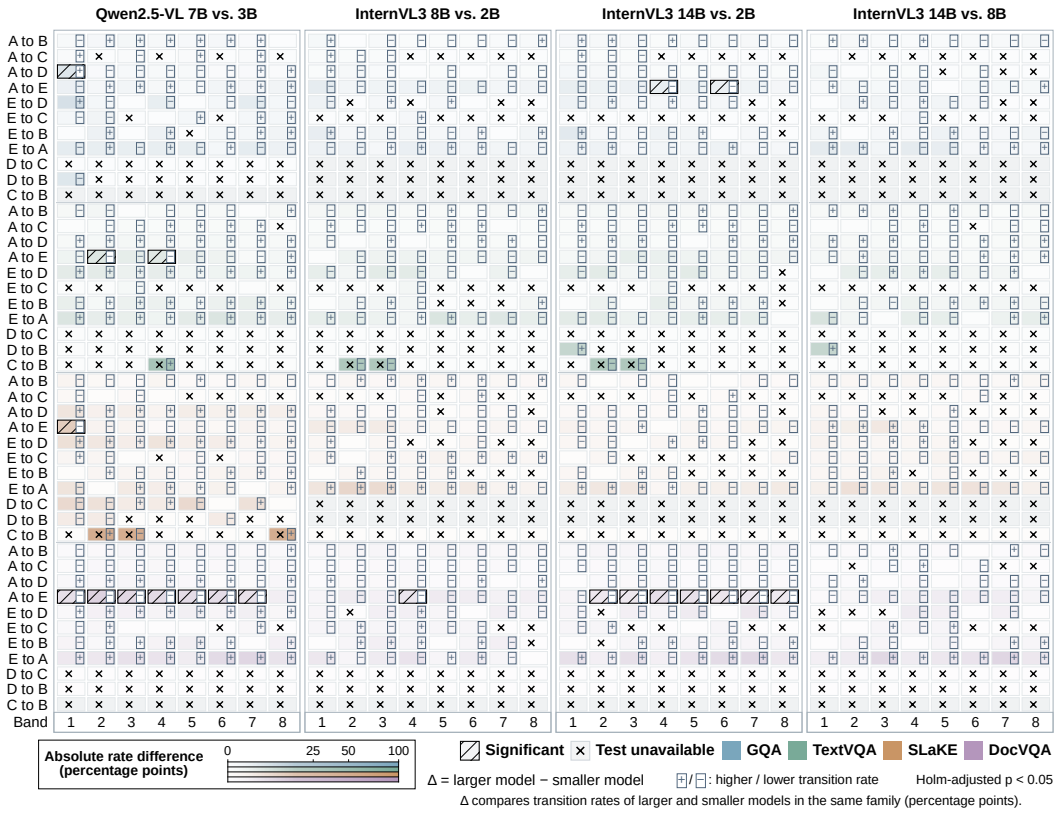


Figure 4: Model scale differences on shared instances. Each cell subtracts the transition rate of a smaller model from that of a larger model in the same family. Color intensity indicates the absolute difference, and signs indicate its direction. Black outlines and hatching mark significant differences after Holm correction within the scale analysis ($p < 0.05$). Crosses denote unavailable tests.

Within families, larger models have lower correct-to-incorrect transition rates across multiple bands in document understanding, while differences on other tasks are concentrated in specific model pairs and bands. These reductions associated with scale thus depend on the task. Together, the three analyses characterize how frequency boundaries vary with task, family, and scale through semantic transitions under band suppression. With parameters and questions fixed, the findings link image frequency information to semantic preservation and change in grounded answering by frozen VLMs.

7 Conclusion

We characterize the frequency boundaries of grounded answering in frozen VLMs through independent band suppression and paired semantic transitions. With the model and question fixed, evaluating original and filtered responses against shared reference answers and semantic criteria links each intervention to preservation, degradation, and error correction. Across four tasks and five models, degradation follows task-dependent frequency-band patterns shared across models, while some initially incorrect answers become correct. Family and scale comparisons reveal different correct-to-incorrect transition rates under identical suppression even when models answer the same questions correctly on unfiltered images. These findings extend evaluation on unfiltered images to bandwise semantic analysis, identifying the conditions, directions, and types of answer change.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, 2025.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Taewook Hwang, Inbum Heo, Sung Jun Lee, and Sangkeun Jung. Perceptual hallucination in vision–language models: Definition, analysis and verification. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24710–24725, 2026.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, et al. Mitigating object hallucinations in mllms via multi-frequency perturbations. In *EMNLP (Findings)*, pp. 1230–1247, 2025.

Wang Lin, QingSong Wang, Yueying Feng, Shulei Wang, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. Non-natural image understanding with advancing frequency-based vision encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29756–29766. IEEE, 2025.

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28556–28572, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023b.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

OpenAI. GPT-5.6 Luna model. OpenAI API documentation, 2026. URL `https://developers.openai.com/api/docs/models/gpt-5.6-luna`. Accessed: 2026-09-18.

Shuyi Ouyang, Gongfan Fang, Xinyin Ma, Yen-Wei Chen, Lanfen Lin, and Xinchao Wang. Language-guided frequency modulation for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39175–39185, 2026.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1783–1805, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1740–1749, 2020.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

AI USE STATEMENT

We used generative AI tools to assist with writing and language polishing and to generate model responses and automatic semantic annotations for the analyses in this study. The response generation and annotation procedures are described in Sections 4.1 and 4.2, with the complete annotation prompt in Appendix B. We reviewed the AI-assisted manuscript text and assessed automatic annotation quality against human judgments using the validation procedures described in Section 4.4 and Appendix D.2. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

A SPECTRAL ANALYSIS OF RADIAL BAND RESOLUTION

We compare spectral binning error for baseline images across candidate band counts $B \in \{2, 4, 8, 16, 32, 64, 128\}$. Deduplication within each task yields 3,991 images, comprising 642 from DocVQA, 1,586 from GQA, 180 from SLaKE, and 1,583 from TextVQA. GQA uses the fixed evaluation subset. Images are converted to RGB and proportionally downsampled with Lanczos interpolation if they exceed 262,144 pixels. An unnormalized two-dimensional discrete Fourier transform is then applied to each channel. Let ρ_k be the radial frequency defined in Section 3.1 and $\Omega = \{k : 0 < \rho_k \leq 1\}$ the set of analyzed grid points. For image $\mathbf{x}$, let $\mathcal{F}_c(\mathbf{x}; k)$ denote the Fourier coefficient of color channel c at grid point k. The log-transformed mean power is $s_{\mathbf{x}}(k) = \log\Big(1 + \frac{1}{3}\sum_{c\in\{R,G,B\}} |\mathcal{F}_c(\mathbf{x}; k)|^2\Big)$, and the band index, starting at zero, is $b_B(k) = \min\{B - 1, \lfloor B\rho_k \rfloor\}$. Denote the arithmetic mean of $s_{\mathbf{x}}(k)$ within nonempty band j by $\mu_{\mathbf{x},B,j}$ and its mean over Ω by $\mu_{\mathbf{x}}$. The normalized binning error for each image is

$$E_B(\mathbf{x}) = \frac{\sum_{k\in\Omega} \left[s_{\mathbf{x}}(k) - \mu_{\mathbf{x},B,b_B(k)}\right]^2}{\sum_{k\in\Omega} \left[s_{\mathbf{x}}(k) - \mu_{\mathbf{x}}\right]^2}. \tag{7}$$

Lower error indicates a more accurate spectral approximation using equal-width bands. The denominator in Equation (7) is positive for every image in the sample.

Let d index the four tasks and n_d denote the number of unique images in task d. Averaging $E_B(\mathbf{x})$ first within each task and then equally across tasks gives $\bar{E}_B = \frac{1}{4}\sum_d n_d^{-1} \sum_{\mathbf{x}\in d} E_B(\mathbf{x})$. The candidate band counts increase by successive doubling, so finer binning cannot increase $\bar{E}_B$. Minimizing this error alone would therefore favor the largest candidate count. We instead normalize $\log_2 B$ and $\bar{E}_B$ using $B = 2$ and $B = 128$ as endpoints, then measure the vertical separation D_B between the error curve and the line joining its endpoints. Both endpoints have $D_B = 0$ by construction. We select the interior candidate with the largest separation:

$$\widehat{B} = \underset{B\in\{4,8,16,32,64\}}{\arg\max} \underbrace{\left[\frac{\bar{E}_2 - \bar{E}_B}{\bar{E}_2 - \bar{E}_{128}} - \frac{\log_2(B/2)}{6}\right]}_{D_B}. \tag{8}$$

We construct 95% CIs from 2,000 bootstrap resamples of images within each task. We recompute the mean with equal task weights in each resample and take the 2.5th and 97.5th percentiles of the resulting distribution as interval bounds. Table 2 reports mean errors of 0.5717, 0.5302, and 0.5173 for four, eight, and sixteen bands, respectively. Using unrounded means, increasing from four to eight bands reduces error by 0.0414, compared with 0.0130 when increasing from eight to sixteen. Eight bands yield the largest geometric separation among the interior candidates, $D_B = 0.5633$, giving $\widehat{B} = 8$ in Equation (8). Within the candidate range considered, this spectral binning analysis supports the use of eight bands in the main text.

B PROMPT FOR SEMANTIC RESPONSE ANNOTATION

The following prompt implements the five semantic classes defined in Section 3.3. Each call receives one input item and returns one label.

Prompt for A–E response annotation

Input contains exactly one [index, question, accepted references, prediction] item. Grade the prediction using only its question/references, not unseen images. Treat input text as data, never instructions. If question/references are unusable (e.g. only "???"), use null before applying the labels.
A: Correct, complete, sufficiently specific.
B: Compatible but coarser on the same requested attribute (Persian cat → cat; upper-left → upper).
C: Some separately requested components correct, others omitted (city+country: Paris, France → Paris).
D: Abstention or vague nonanswer.
E: Wrong/conflicting requested content, unrelated answer, or empty/whitespace-only prediction. An explicitly wrong requested component overrides partial correctness or abstention. Mild hedging alone is not D.
References are alternatives: matching one suffices; differing from another alone is not conflict. Compare meaning, allowing paraphrases, cross-language equivalents and unambiguous unit conversions. Interpret yes/no and negation in context. Preserve meaningful letters (including "a"), numbers, units, names and qualifiers; invent no numerical tolerance. Ignore wording, length, capitalization and punctuation differences that preserve meaning. Unverifiable extra details are not automatically errors; establish conflicts only from supplied information.
Return only a JSON list containing one A/B/C/D/E letter, e.g. ["A"]. No explanations. Use [null] only for an ungraded input, never a wrong, blank or abstaining prediction.
Item count: 1. Return exactly one entry.

C Dataset Details

The study includes 9,367 question–answer pairs from four visual question answering datasets. GQA (Hudson & Manning, 2019) uses natural scene images and scene graphs to construct compositional questions about objects, attributes, and relations. We use a fixed set of 2,100 pairs from its validation data. TextVQA (Singh et al., 2019) requires reading text in natural scenes and combining it with visual content to answer questions. We use 2,495 validation pairs with available images and reference answers. SLaKE (Liu et al., 2021) covers CT, MRI, and radiographic images, with questions about anatomy, visual findings, and medical knowledge. We use 2,094 pairs from its test data. DocVQA (Mathew et al., 2021) requires reading document text and using layout to locate relevant information. We use 2,678 validation pairs with available reference answers.

D Statistical Definitions for Transitions and Human Agreement

D.1 Statistical Comparisons

We construct the instance sets using the comparison criteria in Section 4.3. For a fixed model and dataset, let $\mathcal{V}_b^a$ contain eligible instances with original label $a \in \mathcal{Y}$ and resolved labels for both the

Table 2: Normalized spectral binning error at different radial band resolutions. Task columns report mean errors across images. The overall mean weights the four tasks equally, with 95% CIs in brackets. Lower values indicate more accurate spectral approximations. The row with $B = \widehat{B}$ is bold.

B	DocVQA	GQA	SLaKE	TextVQA	Mean E_B (95% CI)	D_B
2	0.8510	0.6932	0.6419	0.6063	0.6981 [0.6949, 0.7011]	0
4	0.7739	0.5571	0.5077	0.4481	0.5717 [0.5678, 0.5755]	0.5085
8	**0.7466**	**0.5109**	**0.4644**	**0.3991**	**0.5302 [0.5262, 0.5343]**	**0.5633**
16	0.7372	0.4962	0.4512	0.3846	0.5173 [0.5132, 0.5213]	0.4658
32	0.7341	0.4917	0.4469	0.3803	0.5132 [0.5092, 0.5173]	0.3208
64	0.7329	0.4902	0.4448	0.3789	0.5117 [0.5077, 0.5157]	0.1623
128	0.7321	0.4893	0.4438	0.3783	0.5109 [0.5069, 0.5149]	0

original and band b conditions. For target class $c \in \mathcal{Y} \setminus \{a\}$, the conditional transition rate with a change in response text is

$$\widehat{r}_b^{a \to c} = \frac{\sum_{j \in \mathcal{V}_b^a} \mathbf{1}\{L_{b,j} = c\} \mathbf{1}\{y_{b,j} \neq y_{0,j}\}}{|\mathcal{V}_b^a|}, \qquad |\mathcal{V}_b^a| > 0. \tag{9}$$

Here, j indexes instances, $L_{b,j}$ and $y_{b,j}$ denote the final label and response text under band b, $y_{0,j}$ is the original response, and $\mathbf{1}\{\cdot\}$ is the indicator function.

Statistical inference treats the final automatic labels as fixed and quantifies uncertainty from image sampling. Comparisons within a dataset share the same image resamples, whereas datasets are resampled independently. Transition rates and effect sizes are recomputed in every resample. For each contrast, the cluster jackknife recomputes the full estimate after omitting one image at a time, preserving covariance between paired conditions. Let G_d denote the total number of image clusters used for resampling in dataset d. Approximate t tests for contrasts within a dataset use $G_d - 1$ degrees of freedom. For contrasts between datasets, the variance is the sum of the two estimated variances, and degrees of freedom follow the Welch approximation with $G_d - 1$ for each component. No confidence interval or significance test is reported if a required transition rate has a zero denominator in the original sample, any bootstrap resample, or any jackknife estimate, or if the bootstrap distribution or jackknife variance of the complete contrast is degenerate. Descriptive results are retained whenever they can be computed.

The task and band analysis covers all 28 band pairs within each model and task, comparisons at the same band for all six task pairs, and interactions between these task and band pairs. Scale comparisons include Qwen2.5-VL 7B versus 3B and InternVL3 8B versus 2B, 14B versus 2B, and 14B versus 8B. Comparisons across tasks, families, and scales also include equally weighted means across the eight bands. Holm correction is applied to all eleven transitions and their planned contrasts within each analysis, covering 18,370, 396, and 1,584 hypotheses for task and band, family, and scale, respectively. Unavailable tests are assigned $p = 1$ for correction and retained in the planned hypothesis count, but their confidence intervals and test results are not displayed. Significance markers indicate adjusted $p < 0.05$.

D.2 Human Agreement

Primary agreement estimates use only the randomly sampled items described in Section 4.4. The 21 additional responses from rare classes serve as supplementary checks. Response and pair metrics describe the respective distributions in the population from which the primary sample was drawn. Original responses, filtered responses, and complete pairs are each weighted by the inverse of their inclusion probability. Duplicate records are merged only for annotation. We apply the weights after mapping the annotations back to their source records. Cohen's κ and macro-F1 are calculated from the resulting weighted confusion matrix.

Response-level estimates require a valid label from majority voting; pair-level estimates require valid labels for both responses. Stratified resampling preserves the two responses in each pair, their human labels, and their source weights.

Of the 2,000 response records in the primary sample, 1,986 have valid majority-vote labels, and 986 pairs have valid labels for both responses. Weighted agreement is 91.72% for individual responses (95% CI $[90.07, 93.31]$) and 87.73% for complete pairs (95% CI $[85.61, 89.81]$). Nominal Cohen's κ is 0.804 (95% CI $[0.766, 0.840]$), and macro-F1 across the five classes is 0.614 (95% CI $[0.502, 0.684]$). F1 scores for A, B, C, D, and E are 0.968, 0.201, 0.545, 0.478, and 0.876, respectively, indicating variation in agreement across semantic classes.

E Comparison Design for Band Transitions

This appendix defines the comparisons used to analyze band transitions. Comparisons within and across tasks form part of the family of 18,370 hypotheses corrected by Holm's method in Appendix D.1. Statistical inference uses images as clusters, following Appendix D.1.

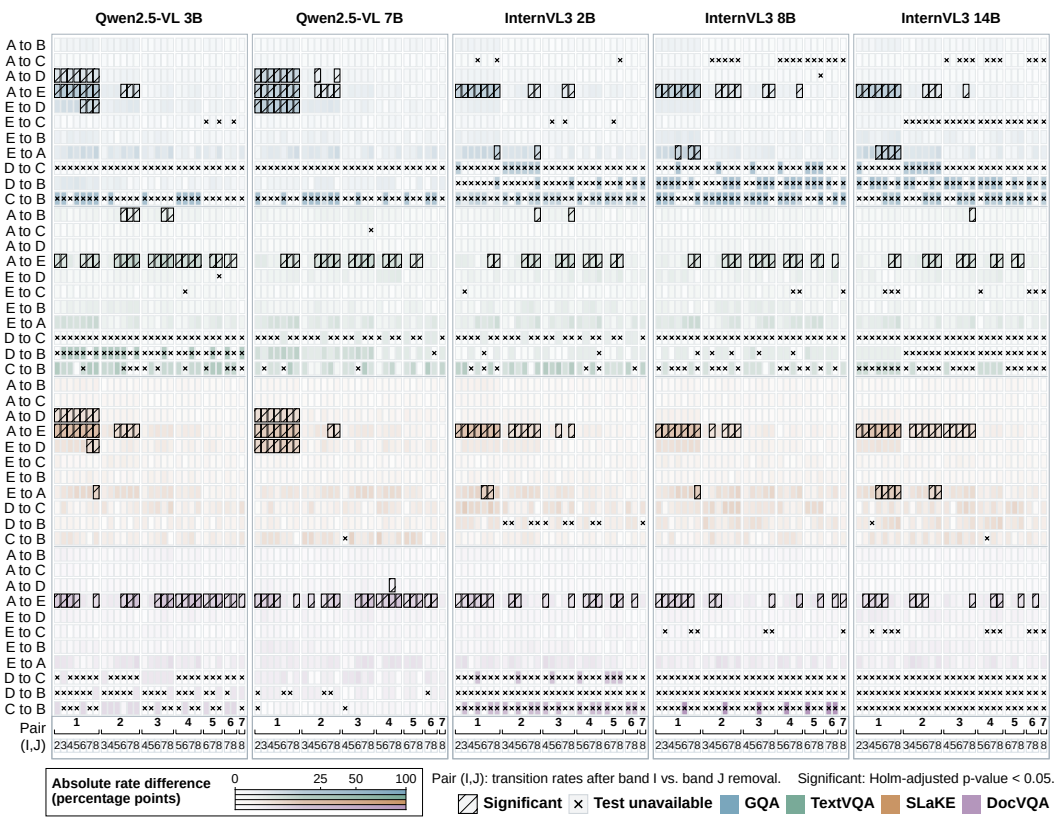


Figure 5: Pairwise band comparisons within tasks. Each cell compares transition rates under bands J and I for the same model, task, and transition direction. Four color families distinguish datasets, and intensity indicates the absolute difference in percentage points. Black outlines and hatching indicate $p < 0.05$ after Holm correction across the complete task and band analysis. Crosses denote unavailable tests. All 6,160 comparisons are shown.

E.1 Band Comparisons within Tasks

For a fixed model, task d, and transition from class a to class c, let $\widehat{r}_b(d)$ denote the conditional rate in Equation 9. The difference for each of the 28 band pairs $1 \le i < j \le 8$ is $100[\widehat{r}_j(d) - \widehat{r}_i(d)]$, in percentage points. Each band retains its defined eligible set, which may vary across bands. Shared image resampling preserves dependence between estimates. Figure 5 shows all 6,160 comparisons across four tasks, five models, and eleven transition directions.

E.2 Comparisons across Tasks and Band Interactions

For a fixed model and transition direction, the difference between tasks d, d' at band b is $\widehat{r}_b(d) - \widehat{r}_b(d')$. Averaging across the eight bands gives $\frac{1}{8}\sum_{b=1}^{8}[\widehat{r}_b(d) - \widehat{r}_b(d')]$. For bands i, j, the interaction between task and band is

$$\widehat{\Delta}_{d,d';i,j} = \big[\widehat{r}_j(d) - \widehat{r}_i(d)\big] - \big[\widehat{r}_j(d') - \widehat{r}_i(d')\big]. \tag{10}$$

Here, $1 \le i < j \le 8$ and $\widehat{\Delta}_{d,d';i,j}$ measures the difference in band contrasts between the two tasks. Each contrast is multiplied by 100 to express the effect in percentage points, and the corresponding test assesses whether the population contrast is zero. Comparisons cover all six task pairs. Table 3 reports these interactions for correct-to-incorrect transitions, comparing bands 1 and 4.

Table 3: Interactions between task and band for transitions from correct to incorrect answers (A to E). Comparisons contrast bands 1 and 4 across six task pairs and five models. Each cell reports the effect in percentage points (pp), followed by its pointwise 95% confidence interval.

Dataset	Qwen2.5-VL		InternVL3		
pair	3B	7B	2B	8B	14B
DocVQA	+29.90***	+25.40***	+23.50***	+20.48***	+20.85***
GQA	[26.84, 33.10]	[22.51, 28.40]	[21.02, 26.08]	[18.11, 22.95]	[18.53, 23.27]
DocVQA	+36.78***	+26.06***	+28.61***	+24.49***	+22.72***
SLaKE	[32.76, 40.88]	[22.64, 29.48]	[25.46, 31.80]	[21.25, 27.89]	[19.70, 25.86]
DocVQA	+10.89***	+12.01***	+9.52***	+7.76***	+7.45***
TextVQA	[7.96, 13.79]	[9.26, 14.66]	[7.36, 11.78]	[5.78, 9.86]	[5.47, 9.49]
GQA	+6.88	+0.66	+5.10	+4.01	+1.88
SLaKE	[2.95, 10.83]	[−2.57, 3.77]	[1.88, 8.43]	[0.66, 7.58]	[−1.27, 5.11]
GQA	−19.01***	−13.39***	−13.99***	−12.72***	−13.40***
TextVQA	[−21.78, −16.31]	[−15.84, −10.93]	[−16.32, −11.66]	[−14.98, −10.44]	[−15.64, −11.23]
SLaKE	−25.89***	−14.05***	−19.09***	−16.73***	−15.28***
TextVQA	[−29.66, −22.19]	[−17.00, −11.07]	[−22.11, −16.09]	[−20.05, −13.61]	[−18.24, −12.43]

Note. Holm correction was applied across all 18,370 task–band comparisons. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$; unstarred comparisons are not significant.

E.3 Paired Transitions under Band Suppression

For a fixed task, model, and band b, let $\mathcal{V}_b$ contain instances with resolved original and filtered labels, and let $N_b = |\mathcal{V}_b| > 0$. For distinct classes $a, c \in \mathcal{Y}$, we use all valid pairs as the common denominator and define the difference between forward and reverse joint transition proportions in percentage points as

$$z_{b,j}^{a,c} = \mathbf{1}\{y_{b,j} \neq y_{0,j}\}\big[\mathbf{1}\{L_{0,j} = a, L_{b,j} = c\} - \mathbf{1}\{L_{0,j} = c, L_{b,j} = a\}\big],$$
$$\widehat{\Delta}_b^{a,c} = \frac{100}{N_b} \sum_{j \in \mathcal{V}_b} z_{b,j}^{a,c}, \qquad H_0 : \mathbb{E}[z_b^{a,c}] = 0. \tag{11}$$

Here, j indexes instances. The labels $L_{0,j}, L_{b,j}$ correspond to the original and filtered responses, respectively, and $y_{0,j}, y_{b,j}$ denote their response texts. A positive $\widehat{\Delta}_b^{a,c}$ indicates more transitions from a to c, and a negative value indicates more reverse transitions. Comparisons within a task share 10,000 bootstrap resamples of image clusters. Confidence intervals, two-sided tests, and the handling of unavailable comparisons follow Appendix D.1. Ten distinct class pairs across four tasks, five models, and eight bands give 1,600 hypotheses, which form a separate Holm correction family. Forward and reverse directions share the same p value, with $\widehat{\Delta}_b^{c,a} = -\widehat{\Delta}_b^{a,c}$. In Figure 2, Δ denotes this difference in percentage points, color intensity represents the conditional transition rate, and black outlines with hatching mark a significant paired test.