Evaluating Vision-Language Models on Bistable Images

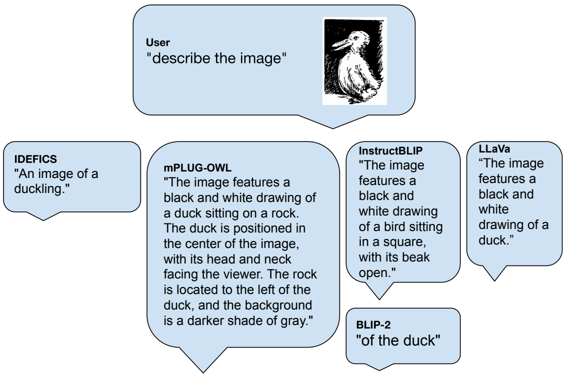
Artemis Panagopoulou[*], **Coby Melkin**[*], **Chris Callison-Burch**
University of Pennsylvania
Correspondence: artemisp@seas.upenn.edu

Abstract

Bistable images, also known as ambiguous or reversible images, present visual stimuli that can be seen in two distinct interpretations, though not simultaneously, by the observer. In this study, we conduct the most extensive examination of vision-language models using bistable images to date. We manually gathered a dataset of 29 bistable images, along with their associated labels, and subjected them to 121 different manipulations in brightness, tint, rotation, and resolution. We evaluated twelve different models in both classification and generative tasks across six model architectures. Our findings reveal that, with the exception of models from the Idefics family and LLaVA1.5-13b, there is a pronounced preference for one interpretation over another among the models, and minimal variance under image manipulations, with few exceptions on image rotations. Additionally, we compared the models' preferences with humans, noting that the models do not exhibit the same continuity biases as humans and often diverge from human initial interpretations. We also investigated the influence of variations in prompts and the use of synonymous labels, discovering that these factors significantly affect model interpretations more than image manipulations showing a higher influence of the language priors on bistable image interpretations compared to image-text training data. All code and data is open sourced [1].

1 Introduction

Bistable images, also known as ambiguous or reversible images, offer unique visual stimuli that present two distinct interpretations, though a viewer cannot simultaneously perceive both (Khalil, 2021). An example of this is depicted in Figure 1, which can be seen as either a rabbit or a duck. The rapid advancements in vision-language models (VLMs) (Ye et al., 2023; Radford et al., 2021; Dai et al., 2023; Liu et al., 2023b; Li et al., 2023a) have sparked interest in testing these models against various types of visual challenges, including optical illusions. While considerable research has been done on how these models interpret geometric and color-varying optical illusions (Guan et al., 2023; Villa et al., 2019; Zhang et al., 2023b; Afifi and Brown, 2019; Benjamin et al., 2019; Sun and Dekel, 2021), exploration into their performance with bistable images remains sparse.



User
"describe the image"

IDEFICS
"An image of a duckling."

mPLUG-OWL
"The image features a black and white drawing of a duck sitting on a rock. The duck is positioned in the center of the image, with its head and neck facing the viewer. The rock is located to the left of the duck, and the background is a darker shade of gray."

InstructBLIP
"The image features a black and white drawing of a bird sitting in a square, with its beak open."

LLaVa
"The image features a black and white drawing of a duck."

BLIP-2
"of the duck"

Figure 1: Depiction of generative models' descriptions of a Duck-Rabbit image. Responses are drawn directly from model outputs.

Motivated by this gap, this work aims to conduct a comprehensive investigation into how vision-language models process and interpret bistable images. We assemble the largest dataset of bistable images to date, apply a range of visual transformations, and examine the models' interpretations and their alignment with human perception.

In particular, we collect 29 bistable images from diverse online sources and cognitive science literature. Each image is subjected to 121 transformations affecting brightness, tint, and resolution resulting in a total of 3,509 processed images. We assessed the behaviors of twelve vision-language

[1] https://github.com/artemisp/Bistable-Illusions-MLLMs.git
[*] authors contributed equally

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 8–29
August 15, 2024 ©2024 Association for Computational Linguistics

models across six distinct model families in both classification and generative settings. Our analysis shows that, apart from a few exceptions, these models generally demonstrate a preference for one interpretation of bistable images over the other. Notably, models from the Idefics family (Laurençon et al., 2024) and LLaVA1.5-13b (Liu et al., 2023b,a) exhibit more balanced preferences. Additionally, while most model responses show little variation to image manipulations, exceptions include CLIP (Radford et al., 2021) and BLIP2 OPT6.7 (Li et al., 2023a), which are sensitive to such changes.

To further understand the influence of training data, we considered multiple models from the same families, trained on identical datasets but using different base language models (LLMs). This approach revealed that even when trained on the same visual data, the models do not consistently align in their preferences, suggesting that LLM priors play a major role in ambiguous image interpretation. This observation underscores that image-text interaction during training is not the sole determinant of how vision-language models perceive ambiguity, echoing earlier findings on the importance of textual signal in VLMs (Jabri et al., 2016; Goyal et al., 2017a; Agrawal et al., 2018).

Additionally, we explored how variations in prompts and the use of synonymous labels affect model interpretations. These textual modifications significantly influenced the models' interpretations, reinforcing the importance of LLM priors on the VLM processing of bistable images. This finding contrasts with previous research on convolutional neural networks (CNNs) focused on geometric optical illusions (Villa et al., 2019; Gomez-Villa et al., 2020; Afifi and Brown, 2019; Benjamin et al., 2019; Sun and Dekel, 2021), which typically show biases consistent with human perception. The CNNs studied did not utilize language model priors, highlighting a fundamental difference in how traditional vision models and VLMs handle visual ambiguity. Our contributions are as follows:

- We have curated the largest collection of bistable images from various online sources and cognitive studies, consisting of 29 unique images. These images have been modified through 121 transformations, creating a comprehensive set of 3.5k images for analysis.
- We analyze the behavior of twelve different vision-language models across six architectural types in both classification and generative tasks, providing a detailed account of their performance on bistable images.
- We examine the influence of prompt variations and synonymous labeling on model interpretations, finding that these textual modifications significantly impact how models perceive bistable images.
- Through direct comparison with human subjects and reference to established cognitive science studies, we assess the degree to which model preferences align with humans. Interestingly, we find that unlike previous work on CNNs (Villa et al., 2019; Gomez-Villa et al., 2020; Afifi and Brown, 2019; Benjamin et al., 2019; Sun and Dekel, 2021), VLMs do not exhibit human biases in bistable images interpretations.

2 Background

2.1 Bistable Images

Bistable images, a unique class of cognitive illusions, present two or more plausible perceptual states, yet viewers cannot observe multiple percepts simultaneously (Khalil, 2021). Instead, observers typically "switch" between the percepts in a seemingly random manner (Kornmeier and Bach, 2005). This phenomenon prompts two primary questions in Cognitive Science regarding bistable images:

1. What causes an individual to initially perceive a particular percept?
2. What triggers the seemingly random switching between percepts?

The exploration of these questions incorporates both bottom-up and top-down considerations (Wang et al., 2013). Bottom-up explanations focus on how the brain processes visual stimuli, starting from the simplest sensory inputs and moving to more complex interpretations. This process involves the detection of subtle visual cues and the neural computation within the visual cortex that ultimately determines the perceived image. Conversely, top-down explanations emphasize the role of cognitive processes, such as expectations, which heavily influence initial perceptions. For instance, a person's previous experiences, like frequently viewing cubes from above, shape their initial interpretation of a Necker Cube (Kuc et al., 2023).

Regarding the switching phenomenon, the dominant bottom-up explanation involves neural mechanisms like spike frequency adaptation or synaptic depression, where the neural connections producing one percept become fatigued, allowing the alternative percept to emerge (Laing and Chow, 2002). Other bottom-up theories propose that this switching is influenced by the brain's inherent noise or randomness (Moreno-Bote et al., 2007) or by unconscious, subtle cues within the images (Ward and Scholl, 2015). On the other hand, top-down explanations suggest that higher cognitive functions, such as motivation and attention, can also induce switching. Studies have shown that individuals can exert some control over their perceptual focus, which influences the switching between different states (Hugrass and Crewther, 2012; Slotnick and Yantis, 2005).

2.2 Vision-Language Models (VLMs)

VLMs integrate visual information as input and generate text as output. VLMs are categorized into contrastive and generative types. Contrastive VLMs, such as the prototypical model CLIP (Radford et al., 2021), are trained to match visual representations with corresponding textual descriptions by distinguishing between different data points. These models create a latent embedding space where similar text and images are drawn closer together, while dissimilar ones are pushed apart. Generative VLMs extend this by incorporating a vision-to-language connection module that projects visual information into the LLM space. This module can either prepend to the input layer of the LLM or condition deeper layers through cross-attention. The integration allows for flexible and dynamic text generation based on visual inputs. For our experiments, we employed models from various families, including CLIP, Idefics (Laurençon et al., 2024), LLaVA1.5 (Liu et al., 2023b,a), mPLUG-Owl (Ye et al., 2023), InstructBLIP (Dai et al., 2023), and BLIP-2 (Li et al., 2023a). Detailed information on the model architectures and the datasets used for training these models is presented in the appendix, in Tables 1 and 2.

2.3 VLMs and Cognitive Illusions

While prior studies have investigated how Convolutional Neural Networks (CNNs) process optical illusions, showing that they often mimic human perceptual errors (Gomez-Villa et al., 2020; Villa et al., 2019; Afifi and Brown, 2019; Benjamin et al., 2019; Sun and Dekel, 2021), the interaction of VLMs with cognitive illusions, especially bistable images, remains underexplored. In contemporary work, Luo et al. (2024) introduce a benchmark designed to evaluate the performance of VLMs on ambiguous, context-dependent visual inputs. Their findings reveal that VLMs significantly underperform compared to humans in these scenarios. More closely related to this work, Zhang et al. (2023b) evaluated VLMs on optical illusions by soliciting binary Yes/No responses and found that larger VLMs tend to be more susceptible to such illusions. However, their study was limited to 16 root images with 100 manually edited variations, focusing primarily on color, shape, and geometric illusions and did not include bistable images. Furthermore, they experimented with only three families of models, whereas our study encompasses six. Limited resources restricted our ability to test some of the larger models that Zhang et al. (2023b) included. Hallusion-bench (Guan et al., 2023) integrates a subset of these optical illusion images, predominantly sourced from Zhang et al. (2023b), but lacks bistable examples.

3 Methodology

3.1 Data Collection

Our dataset comprises 29 bistable images categorized into seven distinct types, sourced from both online platforms, such as Wikipedia, and academic studies (Schooler, 2015; Trautmann, 2021; Wilson, 2012; Pastukhov et al., 2019; Fields et al., 2013; Di Blasi, 2014). Notably, we source all images from the Takashima et al. (2012) research on face perception illusions to compare VLMs to the results of the human study. Among these, twelve images are organized into four classic categories of bistable illusions: the Rubin Vase, Necker Cube, Duck-Rabbit, and Young-Old Woman. Each category includes several iconic versions of the respective illusion type.

To explore the influence of visual modifications on perception, we created 121 variations for each image through a series of controlled manipulations. These manipulations include adjustments to image resolution, rotation, brightness—both increases and decreases—and the application of color tints. The specific colors used for the tints, along with their RGB values, are as follows: red, green, blue, yellow, magenta, and cyan. The intensity of each tint was varied by 0.1 from 0 (no change) to 1.0 (maximum change), and the brightness was adjusted within a range from -1 (darker) to 1 (brighter). We also applied image rotations from 0 to 360 degrees

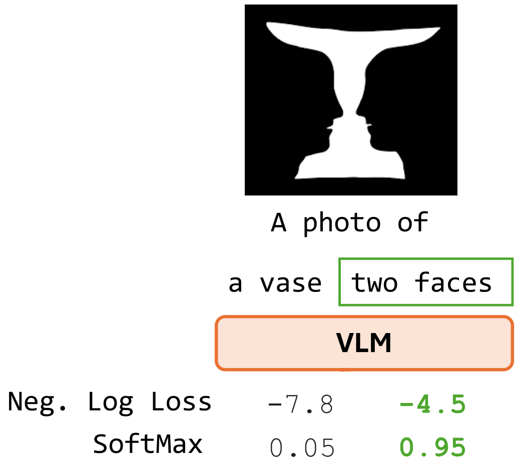


Figure 2: Classification Setup for Generative Models: Each candidate label and corresponding image is forwarded to the model. The prediction is set to be the one with lower loss (higher negative log loss).

every 10 degrees. Finally, we scale the resolution of the images from 0.5 to 1.0 in increments of .1.

3.2 Experimental Setup

We utilized six VLM families, encompassing a total of twelve different models, to evaluate bistable image description. We employed all six VLMs for classification tasks and five for generation tasks (excluding CLIP). The models used and their corresponding implementations on Huggingface Transformers are listed in the footnotes: CLIP (Radford et al., 2021)[2], Idefics 9b (Laurençon et al., 2024)[3], LLaVA1.5 (Liu et al., 2023b,a)[4], mPLUG-Owl (Ye et al., 2023)[5], InstructBLIP (Dai et al., 2023)[6], and BLIP-2 (Li et al., 2023a)[7]. Each model was queried with the default generation parameters and the prompt suggested by their respective model page on Huggingface. All experiments were conducted on a single A100 40GB GPU.

Although all VLMs used, except for CLIP, are generative models, we adapted their outputs to simulate classification. Specifically, we utilized a loss ranking technique (Wei et al., 2022; Li et al., 2021, 2023a; Dai et al., 2023) for classification. As depicted in Figure 2, this technique employs the score to determine the negative log likelihood of each candidate label. In the classification setup, we prompted each VLM with each image along with a pair of strings corresponding to its potential interpretations[8].

In the generative setup, we prompted the models with the format suggested in the HuggingFace documentation for captioning. In addition to model-specific setups, all models were presented with each image and asked to `"describe the image."`

4 Results

4.1 VLMs on Original Images

The models displayed clear preferences between interpretations for the original bistable images. Very rarely were models indifferent between interpretations. The averages between models for our four image categories are shown in Figure 3. We see a strong preference for the 'two faces' interpretation in the Rubin Vase group moderate preferences for 'a cube seen from above' and 'duck' interpretations in Necker Cube and Duck-Rabbit groups. Less classical illusions such at the 'Grimace-Begger, 'Idaho-face', and 'Lion-Gorilla-Tree' also show strong inclinations towards one interpretation. The images with the highest variation across models where the 'Woman-Trumpeter', 'Schroeder Stairs', and 'Raven-Bear' with CLIP variants showing almost consistently opposite preferences to the LLM based generative models.

While the six models generally showed alignment in their interpretation preferences, there was significant variance observed. Figure 4 shows a heat map of model preference correlation coefficients. For more details, refer to Figure 9 in the Appendix which displays the probability distributions for each image category across individual models, revealing some noteworthy model-specific trends. Firstly, all CLIP variants exhibited the exact same probability distributions, with high variance across images within the same category, suggesting a heightened sensitivity to bistability. Secondly, the variants of Idefics 9b and LLaVA 13b demonstrated minimal variance among images of the same category and exhibited relatively moderate preferences, indicating a lower sensitivity to bistability. Moreover, BLIP2-FlanT5, InstructBLIP FlanT5, and mPLUG-Owl showed opposite preferences to CLIP, despite it being used to encode images for these models. This is likely due to the underlying LLM, highlighting the importance of language priors in VLM predictions. Interestingly, all models

[2] openai/clip-vit-base-patch32, openai/clip-vit-base-patch16, laion/CLIP-ViT-B-32-laion2B-s34B-b79K

[3] HuggingFaceM4/idefics-9b, HuggingFaceM4/idefics-9b-instruct

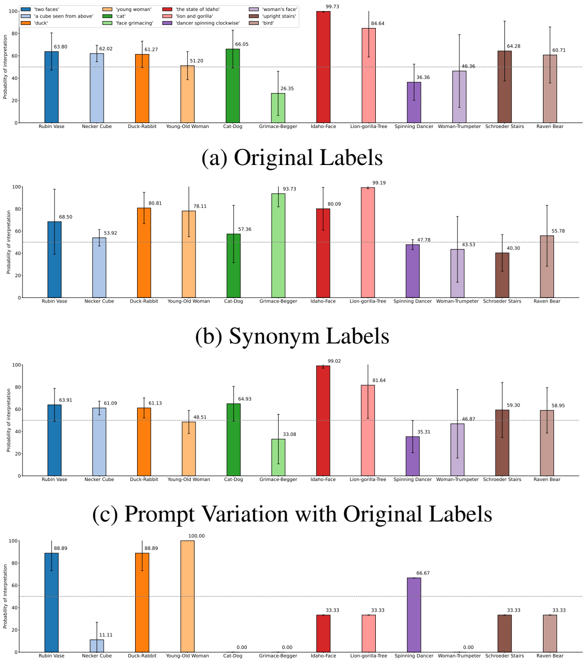
[4] llava-hf/llava-1.5-7b-hf, llava-hf/llava-1.5-13b-hf

[5] MAGAer13/mplug-owl-llama-7b

[6] Salesforce/instructblip-flan-t5-xl

[7] Salesforce/blip2-opt-2.7b, Salesforce/blip2-opt-6.7b, Salesforce/blip2-flan-t5-xl

[8] Image interpretations are found in Appendix C



(a) Original Labels

(b) Synonym Labels

(c) Prompt Variation with Original Labels

(d) Human Initial Interpretations

Figure 3: Between-model averages of probability of the favored interpretation for each image category.

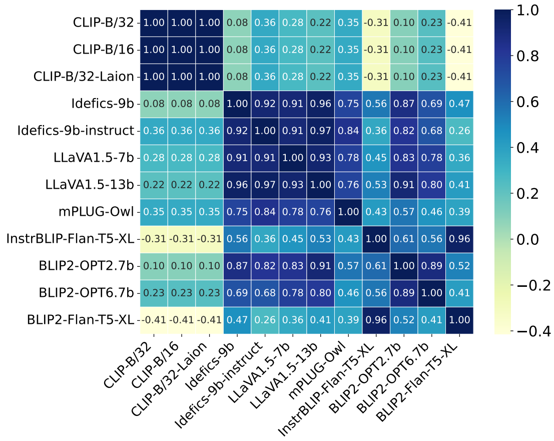


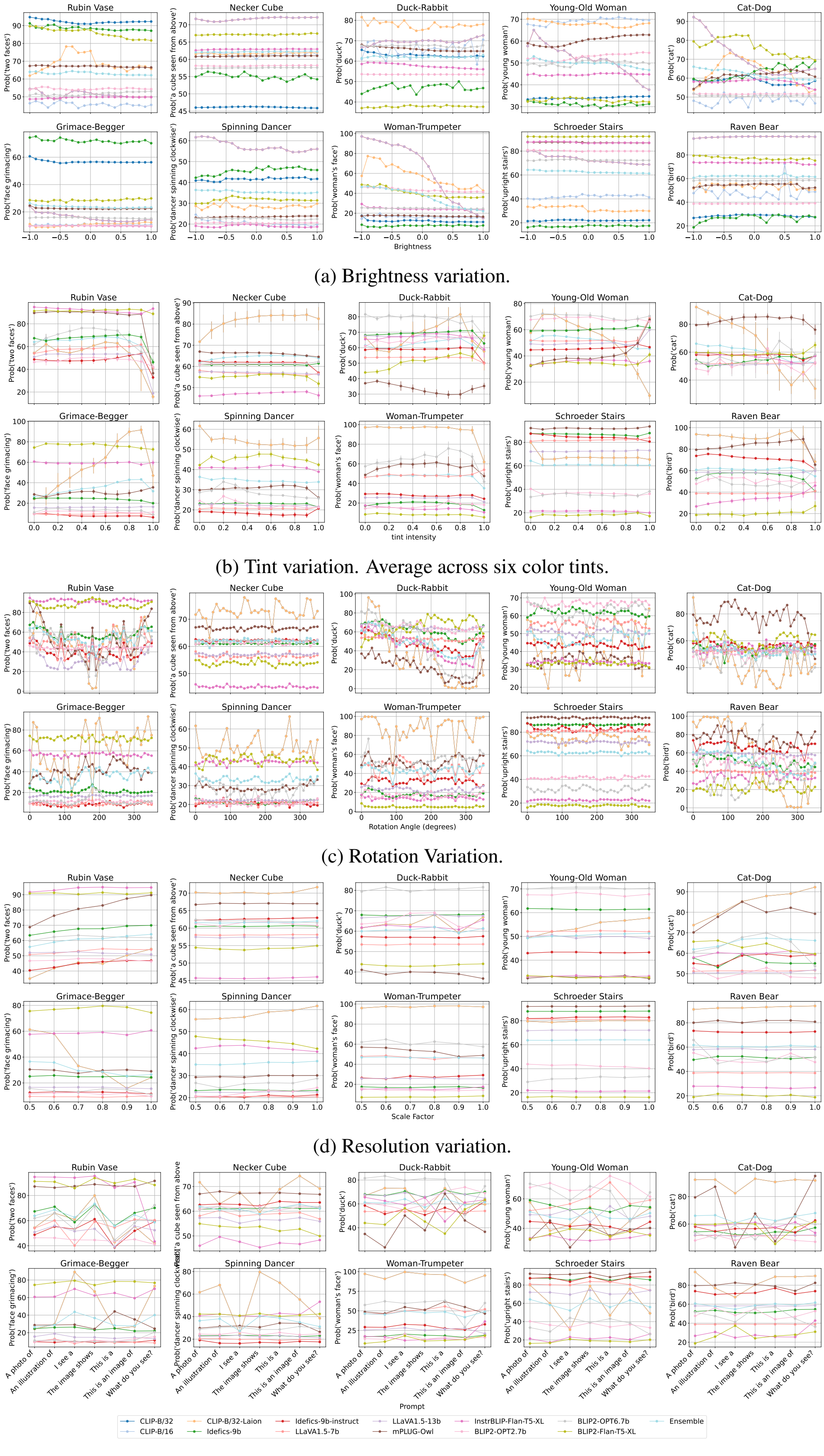
Figure 4: Correlation Among Model Preferences in Original Images

showed a preference for the two animals over the tree in the 'Lion-Gorilla-Tree' illusion, despite the frequent appearance of all these objects in their training sets. Additionally, there was a consistent preference for the face over the full-body abstract silhouette in the 'Grimace-Begger' illusion across all models, except those based on the Flan T5xl architecture. This further accentuates the significant impact of the underlying LLM on image interpretation in VLMs. Notably, although BLIP2 OPT was trained on the same image-text data as the Flan T5 variants, it exhibited almost opposite preferences in some image categories.

4.2 VLMs on Image Manipulations

We observed minimal effects from image manipulations on interpretation probabilities. When adjusting brightness levels, resolution, color tints, and tint intensities, the probabilities for each model remained largely unchanged. Figures 5a and 5b illustrate the minimal impact of these manipulations on model interpretations. This suggests that VLMs tend to overlook minor, low-level perturbations in favor of holistic image processing. Moreover, this finding highlights a significant divergence between VLM processing and human perception of bistable images, which often relies on bottom-up cues according to certain theories (Ward and Scholl, 2015). Notably, the models did not shift interpretations based on subtle cues of brightness and color. The primary exception was the CLIP variants, which demonstrated sensitivity to variations in brightness and tint, particularly in the 'Young-Old Woman,' 'Cat-Dog,' 'Grimace-Begger,' and 'Woman-Trumpeter' illusions. We hypothesize that contrastive learning across aggregation of patches in these models enhance their sensitivity to global changes in the image, as each layer encompasses a more substantial portion of the visual input, making any variations more influential to the model's output. This sensitivity was also observed, though to a lesser extent, in BLIP2-OPT6.7, especially regarding brightness changes in the 'Rubin-Vase' and 'Woman-Trumpeter' illusions. These variations were less pronounced in BLIP2-OPT2.7, particularly for the 'Duck-Rabbit' illusion, and were absent in the corresponding FlanT5-xl variant, underscoring the impact of the underlying LLM's priors on generative vision-language models. Interestingly, when transformations were applied at maximum scale, resulting in a monochrome image, most models exhibited similar preferences, reinforcing the role of language priors in their processing.

Figure 5c shows the variation of interpretations across rotated versions of the images. We find that this manipulation causes significantly higher variation to the color-based manipulations. The variations typically follow the same pattern across models for some bistable images, such as 'Rubin-Vase' and 'Duck-Rabbit'. Notably, contrastive based CLIP-variants once again exhibit the most variation despite being trained with 'minor rotations'



(a) Brightness variation.

(b) Tint variation. Average across six color tints.

(c) Rotation Variation.

(d) Resolution variation.

(e) Prompt variation

Figure 5: Bistable image interpretation under brightness (a), tint (b), rotation (c), resolution (d), and prompt (e) manipulations.

data augmentations. From the generative models mPLUG-Owl seems to exhibit the highest sensitivity to rotation despite also employing rotation augmentation in training. We also observe that the larger LLM variants of LLaVA1.5 and BLIP2-OPT exhibit less variation compared to their smaller counterparts, likely due to the stronger language prior.

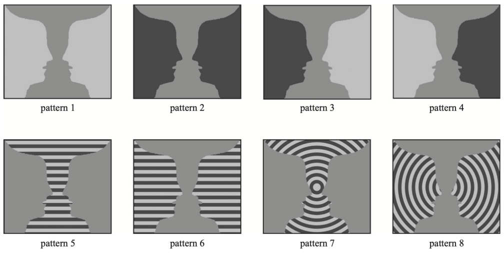


Figure 6: Variations of Rubin Vase images presented to participants in Takashima et al. (2012).

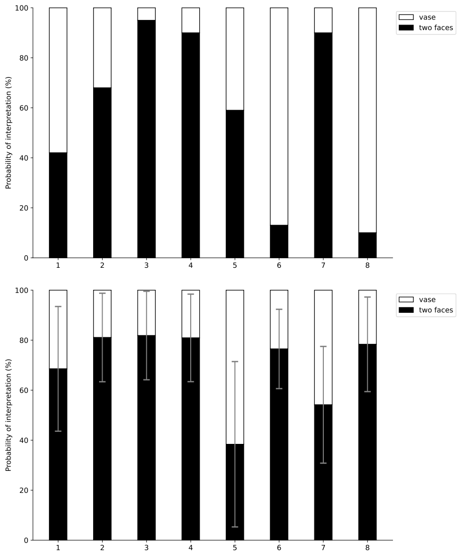


Figure 7: Comparison of between-subject average (top) and between-model average (bottom) probabilities of interpreting each image pattern as two faces in Takashima et al. (2012) and our research, respectively.

4.3 Synonymous Interpretations

To investigate the influence of synonymous interpretation labels on bistable image perception in VLMs, we substituted the original labels with synonyms. Figure 3b displays the effects of these changes on model preferences. The impact is generally mild, but a notable exception occurs with the 'Grimace-Begger' image, where the preference shifts dramatically. In this case, models show a clear preference for interpreting the image as a face rather than a beggar. This shift is likely attributable to the relative unfamiliarity of the synonym 'panhandler' compared to the more commonly recognized term 'face,' making the facial interpretation more likely for the models due to term frequency.

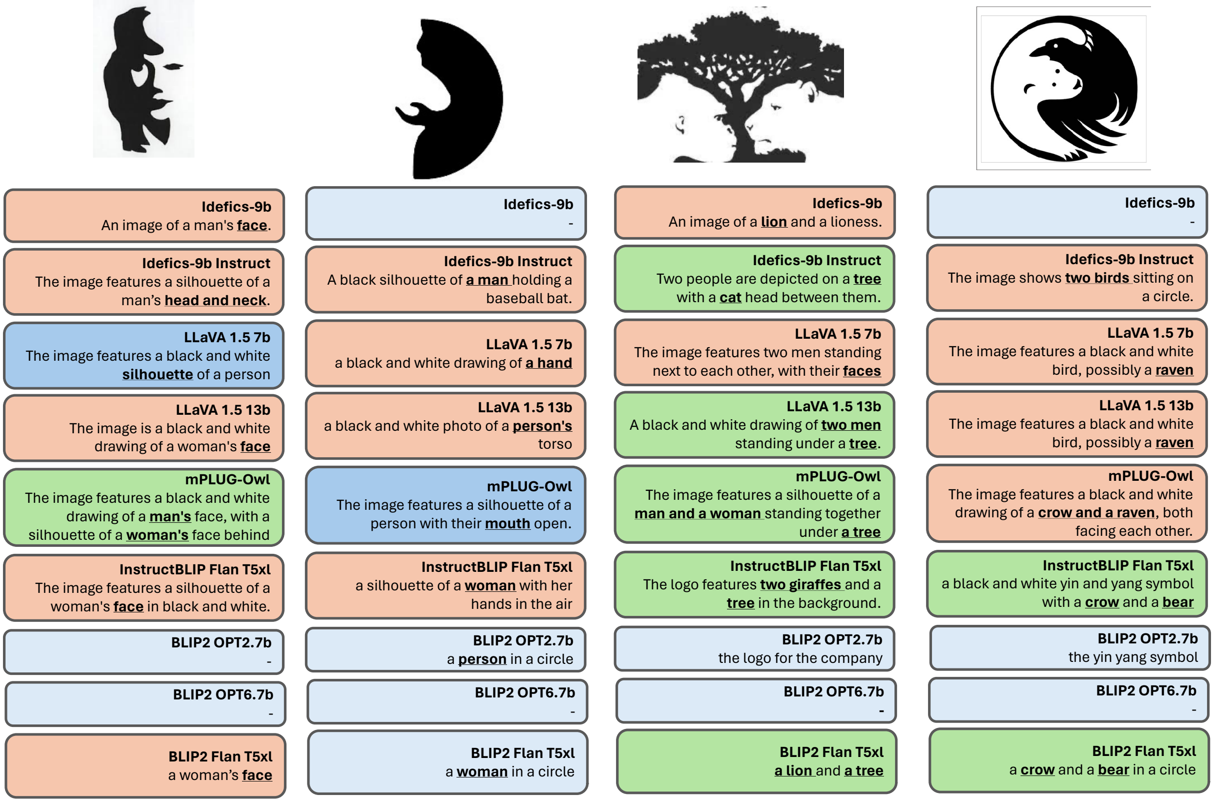
4.4 Prompt Variation

To investigate the effect of prompt variation on VLM bistable image interpretations we examine 7 different prompts. Figure 3c shows little variation on average, however, the individual decomposition of the results in figure 5e shows significant variations within models, especially for CLIP-B/32 and CLIP-B/32-Laion. In fact, while these two models are trained on distinct data of different sizes (400M vs 2B) they exhibit identical behavior across manipulations, indicating the improtance of the architecture in bistable image interpretation. The BLIP family models show higher variation in prompt manipulations compared to LLaVA and Idefics variants. This is likely due to the conditioning of the visual feature extraction module to the instruction prompt.

4.5 Human Interpretations

To compare human initial interpretations with model preferences, we conducted a human evaluation using all original bistable images from our dataset, except for those from Takashima's study (Takashima et al., 2012). We presented these images to three human annotators, asking them to identify "which interpretation they saw first?" They were also given the option to select an alternative interpretation. Figure 3d displays the average results for each interpretation, calculated based on the frequency each interpretation was selected by the annotators across all annotations for that image.

The results reveal a limited correspondence between human and VLM interpretations, contrasting with findings for geometric illusions (Afifi and Brown, 2019; Villa et al., 2019; Gomez-Villa et al., 2020). This discrepancy suggests that the training datasets for VLMs do not trigger the same cognitive biases as those encoded in humans through everyday environmental interactions and conceptual influences. It is important to note that all annotators are students at an American institution, which might influence the results; interpretations could vary significantly based on different socio-cultural experiences and the priors encoded through them.



Idefics-9b
An image of a man's **face**.

Idefics-9b Instruct
The image features a silhouette of a man's **head and neck**.

LLaVA 1.5 7b
The image features a black and white **silhouette** of a person

LLaVA 1.5 13b
The image is a black and white drawing of a woman's **face**

mPLUG-Owl
The image features a black and white drawing of a **man's** face, with a silhouette of a **woman's** face behind

InstructBLIP Flan T5xl
The image features a silhouette of a woman's **face** in black and white.

BLIP2 OPT2.7b
-

BLIP2 OPT6.7b
-

BLIP2 Flan T5xl
a woman's **face**

Idefics-9b
-

Idefics-9b Instruct
A black silhouette of **a man** holding a baseball bat.

LLaVA 1.5 7b
a black and white drawing of **a hand**

LLaVA 1.5 13b
a black and white photo of a **person's** torso

mPLUG-Owl
The image features a silhouette of a person with their **mouth** open.

InstructBLIP Flan T5xl
a silhouette of a **woman** with her hands in the air

BLIP2 OPT2.7b
a **person** in a circle

BLIP2 OPT6.7b
-

BLIP2 Flan T5xl
a **woman** in a circle

Idefics-9b
An image of a **lion** and a lioness.

Idefics-9b Instruct
Two people are depicted on a **tree** with a **cat** head between them.

LLaVA 1.5 7b
The image features two men standing next to each other, with their **faces**

LLaVA 1.5 13b
A black and white drawing of **two men** standing under a **tree**.

mPLUG-Owl
The image features a silhouette of a **man and a woman** standing together under **a tree**

InstructBLIP Flan T5xl
The logo features **two giraffes** and a **tree** in the background.

BLIP2 OPT2.7b
the logo for the company

BLIP2 OPT6.7b
-

BLIP2 Flan T5xl
a lion and **a tree**

Idefics-9b
-

Idefics-9b Instruct
The image shows **two birds** sitting on a circle.

LLaVA 1.5 7b
The image features a black and white bird, possibly a **raven**

LLaVA 1.5 13b
The image features a black and white bird, possibly a **raven**

mPLUG-Owl
The image features a black and white drawing of a **crow and a raven**, both facing each other.

InstructBLIP Flan T5xl
a black and white yin and yang symbol with a **crow** and a **bear**

BLIP2 OPT2.7b
the yin yang symbol

BLIP2 OPT6.7b
-

BLIP2 Flan T5xl
a **crow** and a **bear** in a circle

Figure 8: Depiction of generative models' descriptions for various bistable images. Orange and darker blue colors indicate selection of one interpretation, Green of both, and light blue of neither.

4.6 Replicating Takashima et al. (2012)

We sought to evaluate VLM-human alignment on bistable image processing by comparing our results to a human study. Takashima et al. (2012) presented eight versions of the Rubin Vase illusion n=70 participants. The images are shown in Figure 6 and the human results are shown in Figure 7 (top). They highlight two primary findings: human subjects favored the two faces interpretation for patterns where the profiles' homogeneity is broken (patterns 3 and 4) and favored 'vase' interpretation for patterns where the faces form a continuous background by Gestalt principles (Koffka, 1922) (patterns 6 and 8).

VLMs did not replicate these results, as per the bottom plot in Figure 7. While the models exhibited a strong preference for the 'two faces' interpretation on patterns 3 and 4, the same preference is exhibited in patterns 1 and 2 (where profiles are homogeneous). Furthermore, the models did not exhibit any preference for the 'vase' interpretation in patterns 6 and 8. Even when examined individually in Figure 6 no model exhibited similar patterns to humans. Similar to earlier results, LLaVA and Idefics variants showed high consistency across the images in their tamed preferences. The CLIP variants showed identical patterns despite the varying patch size, unlike in the more global interventions of tint and brightness. Finally, BLIP-2 variants trained on the same image-text data with different LLMs show starkly different preferences, reinforcing the importance of language priors.

4.7 Generative Results

In the generative setup, we performed a qualitative analysis of the results. We found that several interpretation preferences discovered in the classification setup were amplified in generation. Across models, the heavily favored interpretations were faces and ducks for Rubin Vase and Duck-Rabbit images. Figure 1 shows the output of each generative model when prompted to describe a Duck-Rabbit image. Each model employs its own explanatory style, but all favor the duck interpretation. Few models commented on the age of the individual in Young-Old Woman images, but the majority of those comments described the woman as a "girl" or "young woman." An overview of the responses of the models on a subset of the images is delineated in figure 8 and all examples are listed in the Appendix Section D. We observe that most models only comment on a single interpretation, if at all, with some notable exceptions highlighted in green. In few cases, models "hallucinate" descriptions, such as InstructBLIP's interpretation of "two giraffes" for the Lion-Gorilla-Tree illusion. Nevertheless, human inspection of the outputs showed that this was a rare occurrence. We find that for the lion-gorilla-tree image, models are able to identify at least one of the animals, and the tree almost consistently. We hypothesize that this is because of the detail expressed in both interpretations of the image, making it easier even for humans to consciously identify both interpretations simultaneously, even if they are unable to visually perceive both at the same time. Indeed, in the human study, the 'Lion-Gorilla-Tree' image received the most balanced responses across the annotators.

5 Discussion and Limitations

This original analysis of VLM behavior on bistable images has yielded some interesting preliminary results. Similar to humans, VLMs have preferred initial interpretations for most classical bistable images. Five out of six models showed a preference for 'two faces' in Rubin Vase images, 'a cube seen from above' in Necker Cube images, and 'a duck' in Duck-Rabbit images. Young-Old Woman images is the only category for which models' preferences were more neutral and mixed.

We have seen minimal alignment between VLMs and humans when replicating Takashima et al. (2012) and conducting human annotations on the rest of the images. This analysis highlights that VLMs are not sensitive to the same variations that heavily impact human preferences. Models vary greatly in their sensitivity to bistablility. CLIP emerged as a model with strong, variable preferences, while LLaVa is more neutral. CLIP's variability could be attributed to the contrastive pretraining, that might sensitize the model to smaller differences. Moreover, the synthetic nature of bistable images renders them out of domain from most pretraining data, especially for VLMs that are predominantly trained on realistic images.

Nevertheless, making comparisons between human and machine perception of bistable images is difficult beyond the initial biases. Human perception of bistable images exhibits the phenomenon of switching interpretations through extended focus on the image. Replicating the phenomenon of switching is difficult because VLMs take static images at a single point in time. We loosely approximated the movement of time by testing the models on dozens of subtle variations of each image, as discussed above. Under the theories that subtle bottom-up cues precipitate switching in human processing, VLMs do not replicate this phenomenon. We saw that all models' preferences remained steady with variations in brightness, resolution, color, and color tint intensity. Nevertheless, this was in contrast to linguistic variations, highlighting the importance of language priors in generative VLMs.

More research is needed to further our understanding of VLM bistable image interpretation. Using VLMs that process videos could be a tractable way of mimicking the passage of time. Furthermore, additional interventions through design manipulations either through the employment of text-to-image models or human artists could reveal additional insight on VLM behavior for bistable image inputs.

6 Conclusion

In this study we explore the behavior of VLMs on bistable images. We construct the largest bistable image dataset and evaluate 12 different models across six model families under various perturbations: pixel-color based perturbations, resolution, rotations, interpretation label synonyms, and prompt variations. We find that prompts have the highest impact on model preferences whereas, pixel-color perturbations have minimal effects. We further conduct human study comparisons, and find that VLMs do not exhibit the same initial biases on bistable images as human subjects.

Acknowledgments

This research was supported by a gift from AWS AI for research in Trustworthy AI.

References

Mahmoud Afifi and Michael S. Brown. 2019. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 243–252. IEEE.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer*

Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 4971–4980. IEEE Computer Society.

Ari Benjamin, Cheng Qiu, Ling-Qi Zhang, Konrad Kording, and Alan Stocker. 2019. Shared visual illusions between humans and artificial neural networks. In *2019 Conference on Cognitive Computational Neuroscience*, volume 10, pages 2019–1299. Cognitive Computational Neuroscience.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning web-scale training datasets is practical. *ArXiv preprint*, abs/2302.10149.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: web-curated image-text data created by the people, for the people. *Preprint*, arXiv:2111.11431.

Luca Di Blasi. 2014. *Splitting Images: Understanding Irreversible Fractures through Aspect Change*, pages 67–87.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

R Fields, Alfonso Araque, Heidi Johansen-Berg, Soo-Siang Lim, Gary Lynch, Klaus-Armin Nave, Maiken Nedergaard, Ray Perez, Terrence Sejnowski, and Hiroaki Wake. 2013. Glial biology in learning and cognition. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 20.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo. 2020. Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. *Vision Research*, 176:156–174.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *ArXiv preprint*, abs/2310.14566.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Laila Hugrass and David Crewther. 2012. Willpower and conscious percept: volitional switching in binocular rivalry. *PloS one*, 7(4):e35963.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021,* *Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.

Elias L Khalil. 2021. Why does rubin's vase differ radically from optical illusions? framing effects contra cognitive illusions. *Frontiers in Psychology*, 12:597758.

Kurt Koffka. 1922. Perception: an introduction to the gestalt-theorie. *Psychological bulletin*, 19(10):531.

Jürgen Kornmeier and Michael Bach. 2005. The necker cube—an ambiguous figure disambiguated in early visual processing. *Vision Research*, 45(8):955–960.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Alexander Kuc, Vladimir Maksimenko, Andrey Savosenkov, Nikita Grigorev, Vadim Grubov, Artem Badarin, Victor Kazantsev, Susanna Gordleeva, and Alexander Hramov. 2023. Studying perceptual bias in favor of the from-above necker cube perspective in a goal-directed behavior. *Frontiers in Psychology*, 14:1160605.

Carlo R Laing and Carson C Chow. 2002. A spiking neuron model for binocular rivalry. *Journal of computational neuroscience*, 12:39–53.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023c. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv preprint*, abs/2306.04387.

Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, et al. 2024. Codis: Benchmarking context-dependent visual comprehension for multimodal large language models. *ArXiv preprint*, abs/2402.13607.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

Rubén Moreno-Bote, John Rinzel, and Nava Rubin. 2007. Noise-induced alternations in an attractor network model of perceptual bistability. *Journal of neurophysiology*, 98(3):1125–1139.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151.

Alexander Pastukhov, Philipp Kastrup, Isabel Abs, and Claus-Christian Carbon. 2019. Switch rates for orthogonally oriented kinetic-depth displays are correlated across observers. *Journal of Vision*, 19:1.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Jonathan Schooler. 2015. Bridging the objective/subjective divide towards a meta-perspective of science and experience. *In T. Metzinger J. M. Windt (Eds). Open MIND: 34(T). Frankfurt am Main: MIND Group*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv preprint*, abs/2111.02114.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Scott D Slotnick and Steven Yantis. 2005. Common neural substrates for the control and effects of visual attention and perceptual bistability. *Cognitive Brain Research*, 24(1):97–108.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *ArXiv preprint*, abs/2103.01913.

Eric D Sun and Ron Dekel. 2021. Imagenet-trained deep neural networks exhibit illusion-like response to the scintillating grid. *Journal of Vision*, 21(11):15–15.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *ArXiv preprint*, abs/2303.15389.

Midori Takashima, Teruo Fujii, and Ken Shiina. 2012. Face or vase? areal homogeneity effect. *Perception*, 41(11):1392–1394.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Laura Trautmann. 2021. Emotions evoked by geometric patterns. *J*, 4:376–393.

Alexander Gómez Villa, Adrián Martín, Javier Vazquez-Corral, and Marcelo Bertalmío. 2019. Convolutional neural networks can be deceived by visual illusions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12309–12317. Computer Vision Foundation / IEEE.

Megan Wang, Daniel Arteaga, and Biyu J He. 2013. Brain mechanisms for simple perception and bistable perception. *Proceedings of the National Academy of Sciences*, 110(35):E3350–E3359.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *ArXiv preprint*, abs/2401.06805.

Emily J Ward and Brian J Scholl. 2015. Stochastic or systematic? seemingly random perceptual switching in bistable events triggered by transient unconscious cues. *Journal of Experimental Psychology: Human Perception and Performance*, 41(4):929.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Amanda Wilson. 2012. Multistable perception of art-science imagery. *Leonardo*, 45.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv preprint*, abs/2304.14178.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023a. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *ArXiv preprint*, abs/2306.17107.

Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. 2023b. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? *Preprint*, arXiv:2311.00047.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023. Svit: Scaling up visual instruction tuning. *ArXiv preprint*, abs/2307.04087.

A Model Details

We summarize the architectural differences for the models used in our study in Table 1 and list the various datasets they were trained on both for pre-training and instruction tuning (where applicable) in Table 2.

B Additional Results

B.1 Individual Results: Original Images

Figure 9 we list the individual model results for the original labels.

B.2 Individual Results: Synonymous Interpretations

Figure 10 we list the individual model results for the synononymous labels. We find that there is non-trivial variation that is attributed to the likelihood of the terms used as the labels.

B.3 Individual Results: Takashima et al. (2012)

Figure 11 lists the individual results for each model on the Takashima et al. (2012) image study.

B.4 Tint Variation Individual Plots

Figures 12, 13, 14, 15, 16, 17 show the individual variations of each model for each image category based on tint variations. We find limited effect in preferences with highest variability observed by the CLIP variants. Interestingly, most models seem to show same preferences when full tint is applied, indicating a monochrome image - hence the linguistic priors play a large role in model behavior as indicated by the synonym and prompt variation experiments.

C Bistable Image Collection

We present examples of original images in our dataset, without any visual manipulations in figures 18, 19, 20, 21.

D Generative Examples

We present examples of generations from the models prompted with "describe the image" in figures 22, 23, 24, 25, 26 with the exception of few question based prompts: “What is the orientation of the staircase/cube?" for the Shroeder stairs and Necker Cube illusions, and “What is the dancer’s spinning direction?" for ‘Spinning Dancer’.

Model	#Train Param.	LLM	Res.	ViT	LLM Size	V-L Type	V-L Size	#Tokens	Deep V-L	Frozen LLM	Frozen ViT
Idefics 9b (Laurençon et al., 2024)	9b	LLaMA(Touvron et al., 2023)	224	OpenCLIP-H[9]	7b	Perceiver (Jaegle et al., 2021)	194M	64	✓	✓	✓
Idefics 9b Instruct	9b	LlaMA	224	OpenCLIP	7b	Perceiver	194M	64	✓	✓	✓
LLaVA-1.5 7b (Liu et al., 2023b,a)	20M	Vicuna1.5-7B (Chiang et al., 2023)	336	CLIP ViT-L (Radford et al., 2021)	7b	Linear	20M	577	×	×	✓
LLaVA-1.5 13b	20M	Vicuna1.5-13B	336	CLIP ViT-L	13b	Linear	20M	577	×	×	✓
BLIP-2 OPT2.7b (Li et al., 2023a)	188M	OPT2.7b (Zhang et al., 2022)	224	EVA-CLIP-g	2.7b	Q-Former	188M	32	×	✓	✓
BLIP-2 OPT6.7b	188M	OPT6.7b	224	EVA-CLIP-g	6.7b	Q-Former	188M	32	×	✓	✓
BLIP-2 FlanT5xl	188M	FlanT5xl	224	EVA-CLIP-g	3b	Q-Former	188M	32	×	✓	✓
InstructBLIP FlanT5xl (Dai et al., 2023)	188M	FlanT5xl (Chung et al., 2024)	224	EVA-CLIP-g(Sun et al., 2023)	3b	Q-Former (Li et al., 2023b)	188M	32	×	✓	✓
mPLUG-Owl (Ye et al., 2023)	500M	LLaMA	224	CLIP ViT-L	7b	Visual Abstractor (Ye et al., 2023)		64	×	×	×

Table 1: Overview of Generative VLMs architectures examined on their perception of bistable images.

Model	Pretraining Data	Instruction Tuning Data
CLIP	400M image-caption data (undisclosed)	N/A
Idefics 9b	OBELICS (Laurençon et al., 2024), Wikipedia[10],Conceptual Captions(Sharma et al., 2018), Conceptual Captions 12M (Changpinyo et al., 2021), WIT (Srinivasan et al., 2021), Localized Narratives (Pont-Tuset et al., 2020), RedCaps (Desai et al., 2021), COCO (Chen et al., 2015), SBU Captions (Ordonez et al., 2011), Visual Genome (Krishna et al., 2017), YFCC100M (Thomee et al., 2016)	N/A
Idefics 9b Instruct	OBELICS (Laurençon et al., 2024), Wikipedia[11] ,CC3M(Sharma et al., 2018), CC12M (Changpinyo et al., 2021), WIT (Srinivasan et al., 2021), Localized Narratives (Pont-Tuset et al., 2020), RedCaps (Desai et al., 2021), COCO (Chen et al., 2015), SBU (Ordonez et al., 2011), Visual Genome (Krishna et al., 2017), YFCC100M (Thomee et al., 2016)	M3IT (Li et al., 2023c), LRV-Instruction (), LLaVA150k (Liu et al., 2023b),LLaVAR-Instruct (Zhang et al., 2023a),SVIT (Zhao et al., 2023), UltraChat (Ding et al., 2023)
LLaVA-1.5	LLaVA (Liu et al., 2023a) [subsets of LAION-400M (Schuhmann et al., 2021), CC3M (Sharma et al., 2018), SBU (Ordonez et al., 2011)]	VQAv2 (Goyal et al., 2017b), GQA (Hudson and Manning, 2019),OKVQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022),OCRVQA (Mishra et al., 2019), TextCaps (Sidorov et al., 2020), LLaVA150k (Liu et al., 2023b), ShareGPT [12]
BLIP-2	COCO (Chen et al., 2015), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), LAION400M (Schuhmann et al., 2021), Visual Genome (Krishna et al., 2017)	N/A
InstructBLIP	COCO (Chen et al., 2015), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), LAION400M (Schuhmann et al., 2021), Visual Genome (Krishna et al., 2017)	COCO (Chen et al., 2015), Web CapFilt (Li et al., 2023a), TextCaps (Sidorov et al., 2020), VQAv2 (Goyal et al., 2017b), OKVQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), LLaVA150k (Liu et al., 2023b), OCRVQA (Mishra et al., 2019)
mPLUG-Owl	LAION-400M (Schuhmann et al., 2021), COYO (Carlini et al., 2023), COCO (Chen et al., 2015), Laion-en (Schuhmann et al., 2022), DataComp (Gadre et al., 2024)	VQAv2 (Goyal et al., 2017b), OKVQA (Marino et al., 2019), OCR-VQA (Mishra et al., 2019), GQA (Hudson and Manning, 2019), A-OKVQA (Schwenk et al., 2022), RefCOCO (Yu et al., 2016), Visual Genome (Krishna et al., 2017), LLaVA150K (Liu et al., 2023b), ShareGPT, SlimOrca (Lian et al., 2023)

Table 2: Overview of Pretraining and Instruction Tuning Datasets (adapted from Wang et al. (2024))

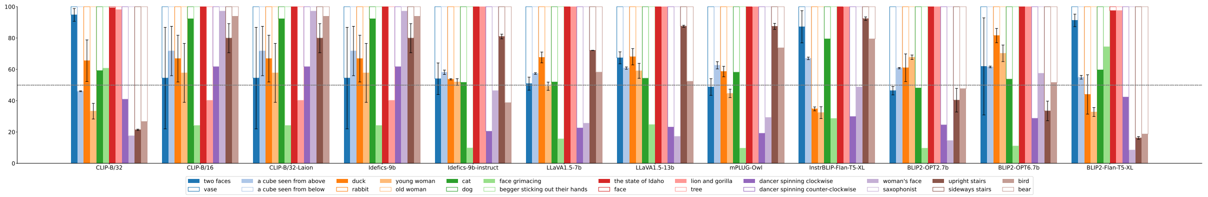


Figure 9: Average probability distributions for each model evaluated on each image category.

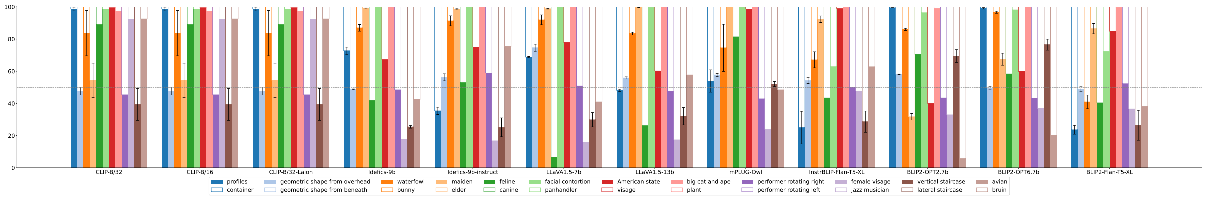


Figure 10: Synonymous Interpretations

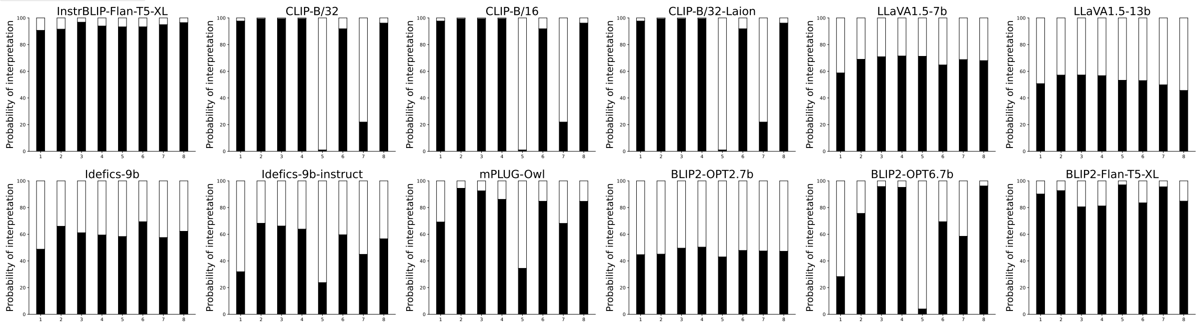


Figure 11: Individual model preferences for Takashima et al. (2012) images.

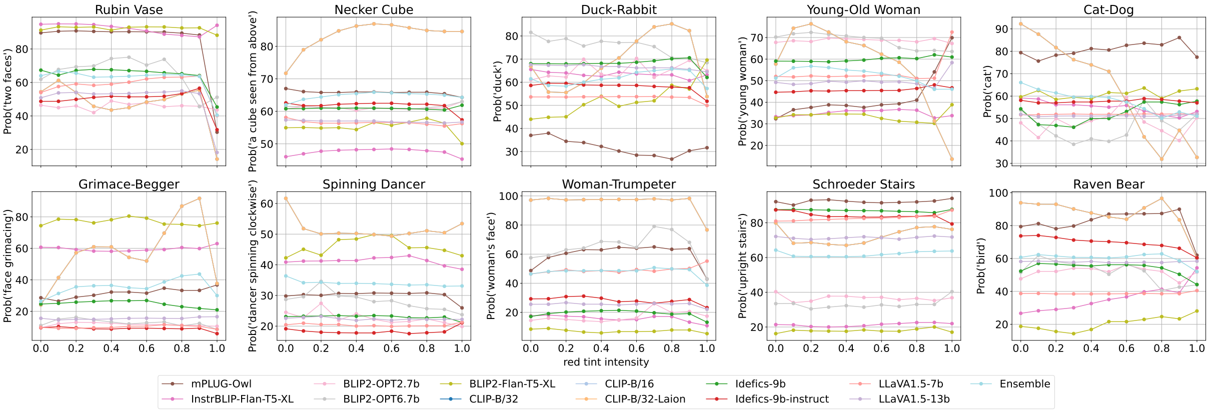


Figure 12: Red Tint Variation

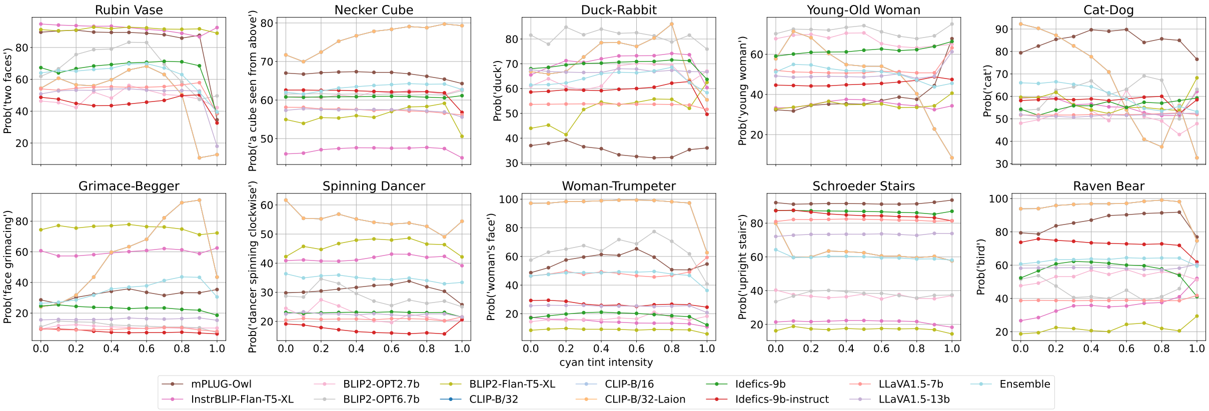


Figure 13: Cyan Tint Variation

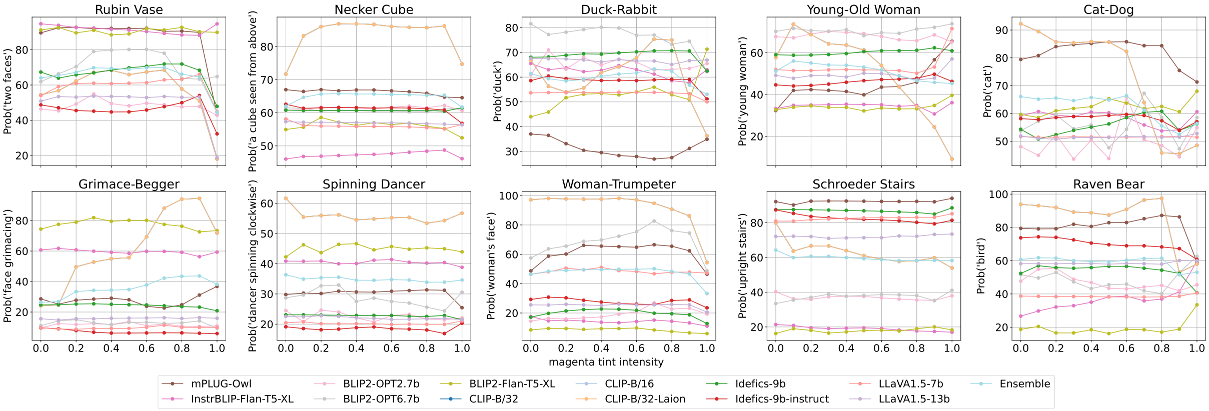


Figure 14: Magenta Tint Variation

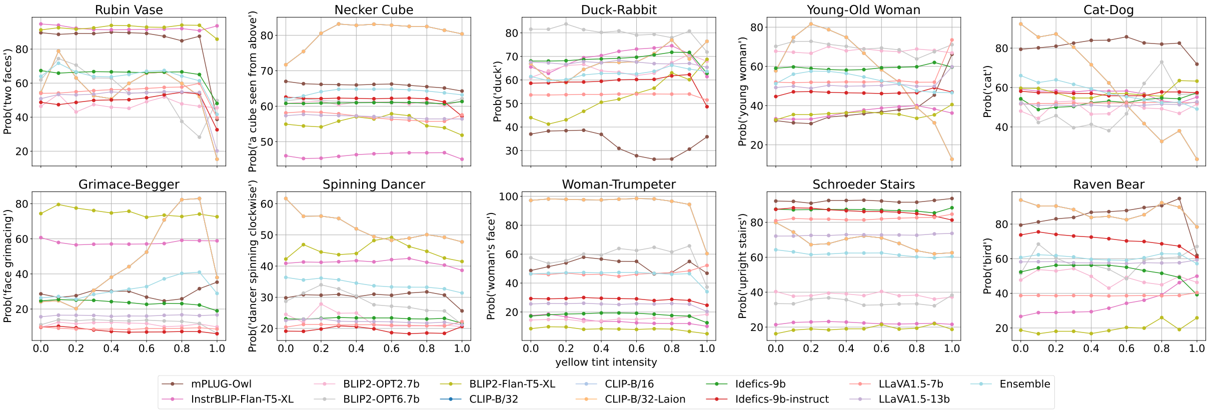


Figure 15: Yellow Tint Variation

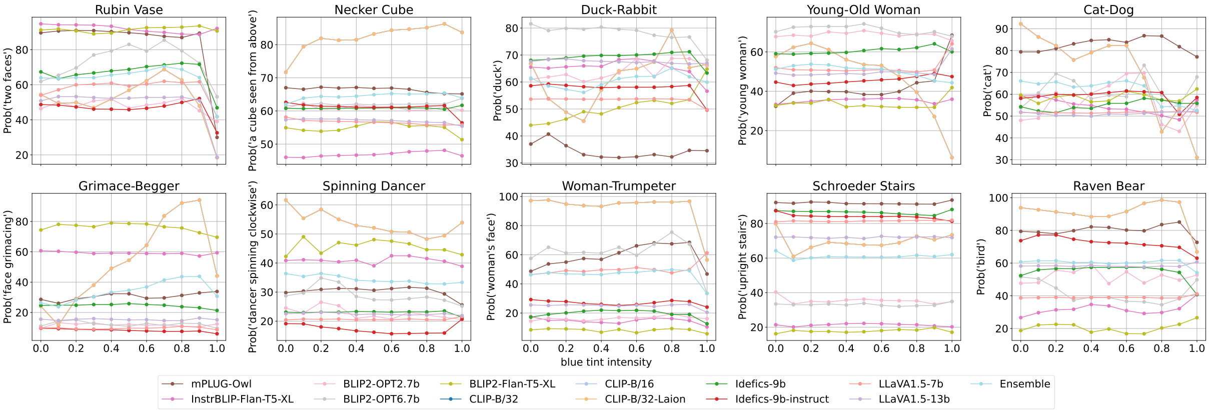


Figure 16: Blue Tint Variation

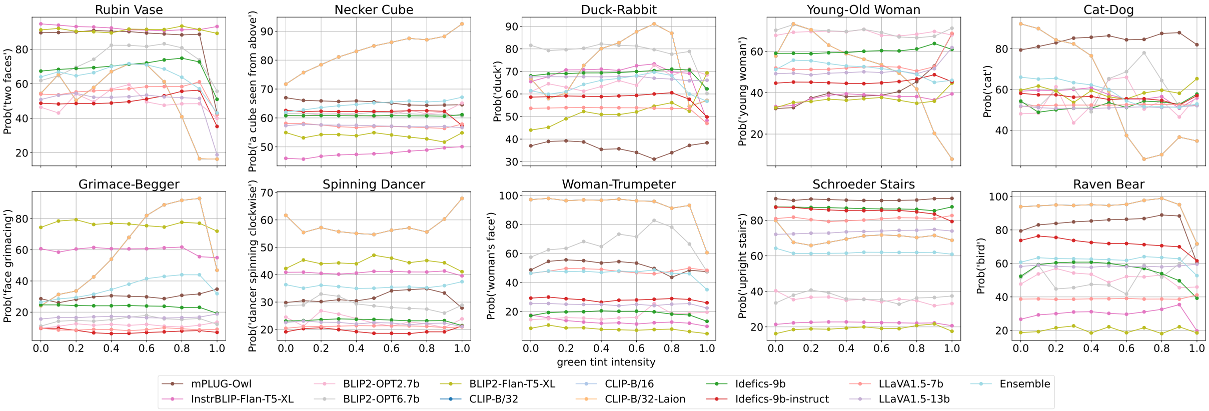


Figure 17: Green Tint Variation

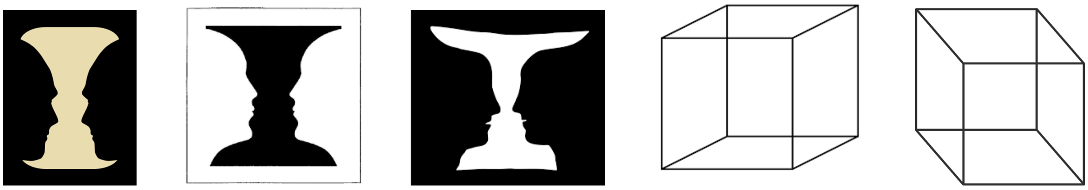


Figure 18: Rubin Vase illusions (interpretations: ["vase", "two faces"]) and Necker Cube illusions (interpretations: ["a cube seen from below", "a cube seen from above"]).

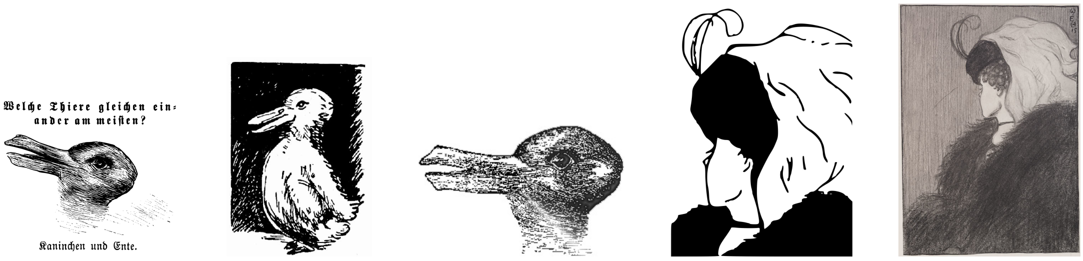


Figure 19: Duck-Rabbit illusion (interpretations: ["duck", "rabbit"]) and Young-Old Woman illusion (interpretations: ["young woman", "old woman"]).

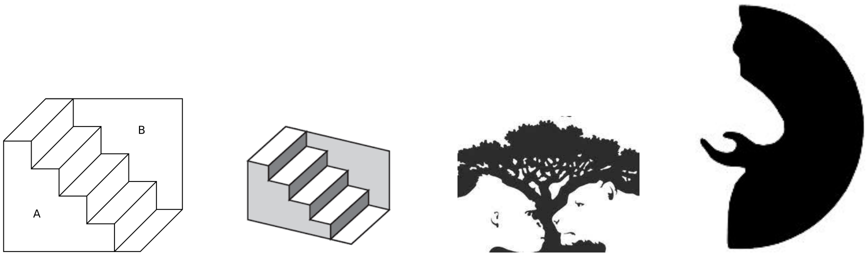


Figure 20: Shroeder Stairs illusion (interpretations: ["upright stairs", "sideways stairs"]), Lion-Gorilla-Tree illusion (interpretations: ["lion and gorilla", "tree"]) and Grimace-Begger illusion (interpretations: ["grimace", "beggar"]).

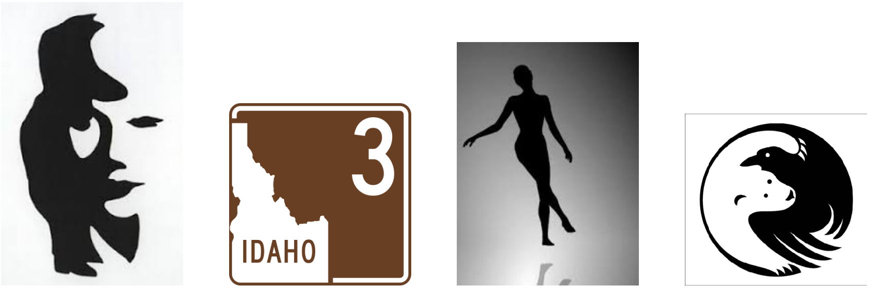
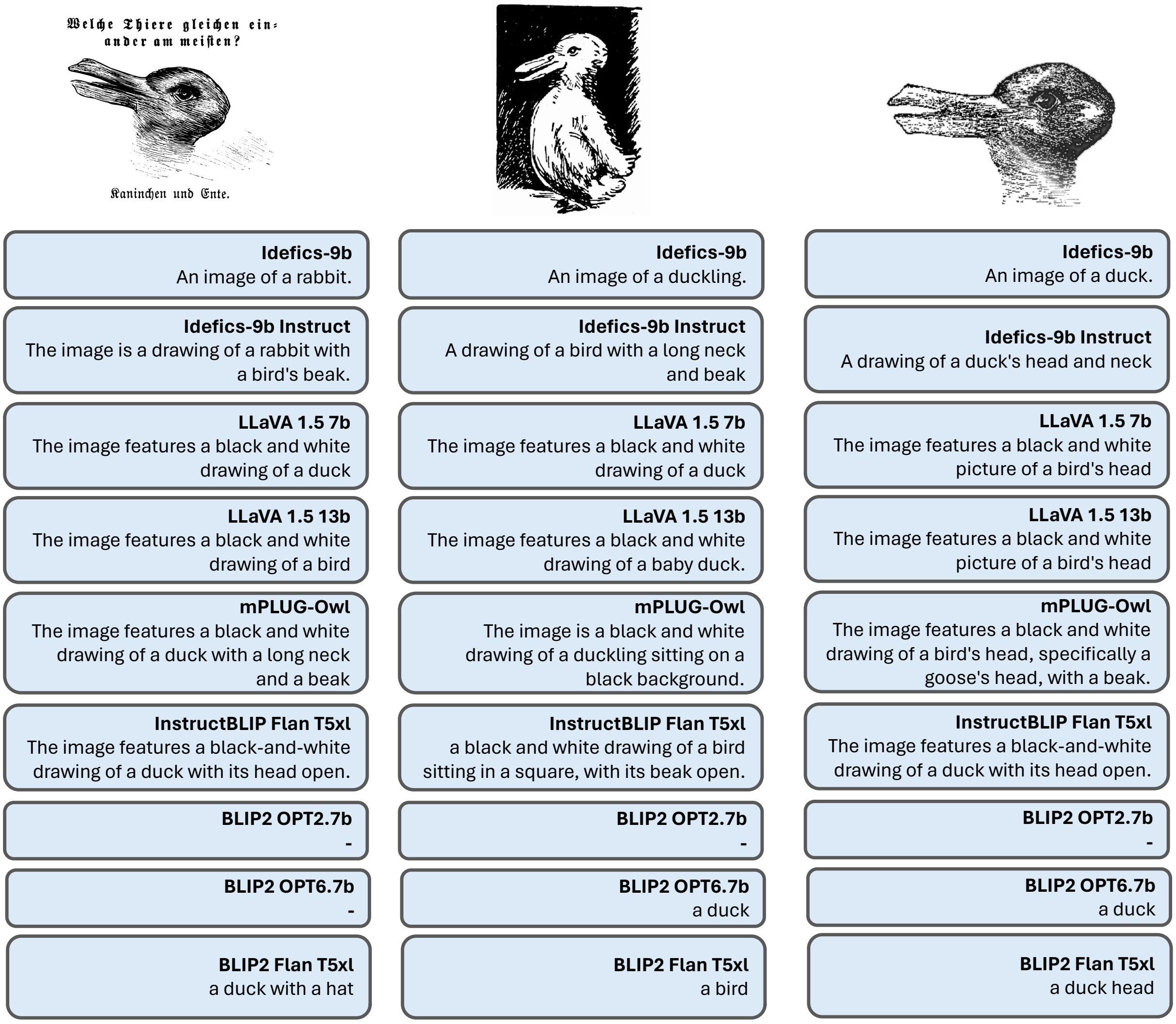
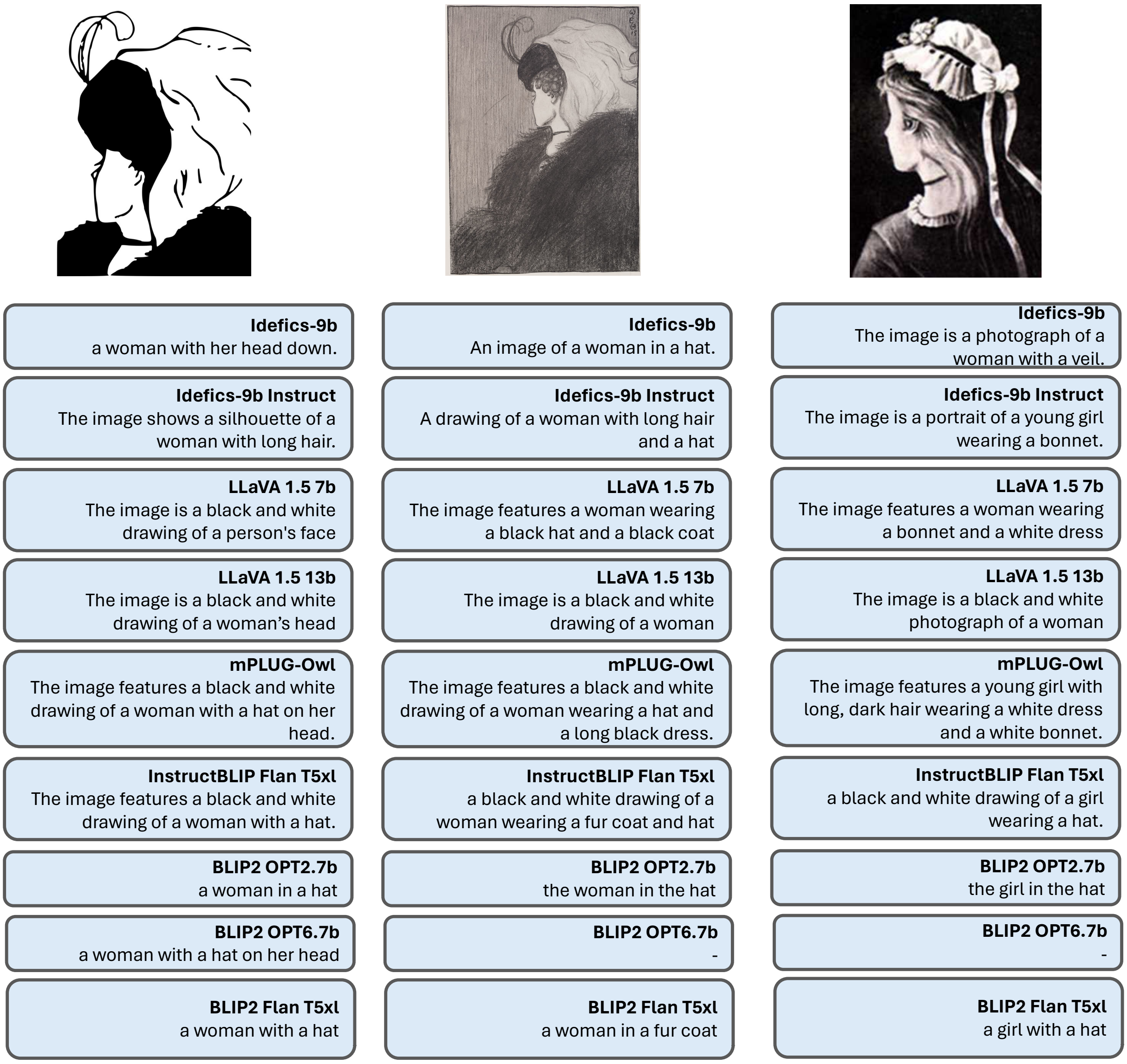


Figure 21: Various illusions from left to right: Woman-Trumpeter (interpretations: ["woman's face", "saxophonist"], Idaho-Face (interpretations: ["the state of Idaho", "face"]), Spinning Dancer (interpretations: ["dancer spinning clockwise", "dancer spinning counter-clockwise"]), and Raven-Bear (interpretations: ["bird", "bear"]).



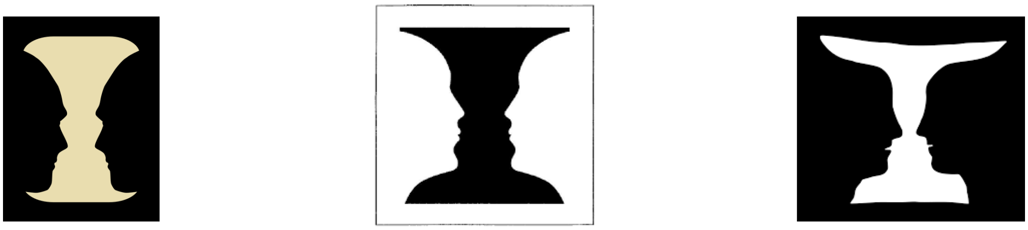
Idefics-9b An image of a rabbit.	**Idefics-9b** An image of a duckling.	**Idefics-9b** An image of a duck.
Idefics-9b Instruct The image is a drawing of a rabbit with a bird's beak.	**Idefics-9b Instruct** A drawing of a bird with a long neck and beak	**Idefics-9b Instruct** A drawing of a duck's head and neck
LLaVA 1.5 7b The image features a black and white drawing of a duck	**LLaVA 1.5 7b** The image features a black and white drawing of a duck	**LLaVA 1.5 7b** The image features a black and white picture of a bird's head
LLaVA 1.5 13b The image features a black and white drawing of a bird	**LLaVA 1.5 13b** The image features a black and white drawing of a baby duck.	**LLaVA 1.5 13b** The image features a black and white picture of a bird's head
mPLUG-Owl The image features a black and white drawing of a duck with a long neck and a beak	**mPLUG-Owl** The image is a black and white drawing of a duckling sitting on a black background.	**mPLUG-Owl** The image features a black and white drawing of a bird's head, specifically a goose's head, with a beak.
InstructBLIP Flan T5xl The image features a black-and-white drawing of a duck with its head open.	**InstructBLIP Flan T5xl** a black and white drawing of a bird sitting in a square, with its beak open.	**InstructBLIP Flan T5xl** The image features a black-and-white drawing of a duck with its head open.
BLIP2 OPT2.7b -	**BLIP2 OPT2.7b** -	**BLIP2 OPT2.7b** -
BLIP2 OPT6.7b -	**BLIP2 OPT6.7b** a duck	**BLIP2 OPT6.7b** a duck
BLIP2 Flan T5xl a duck with a hat	**BLIP2 Flan T5xl** a bird	**BLIP2 Flan T5xl** a duck head

Figure 22: Duck-Rabbit generative examples



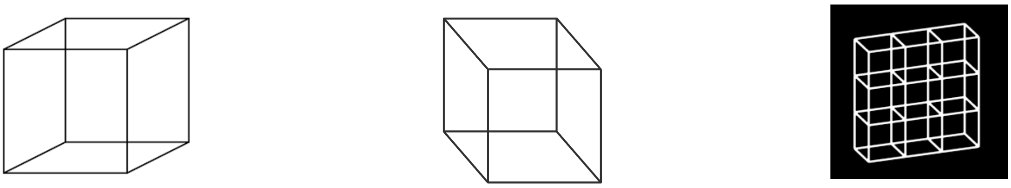
Idefics-9b	a woman with her head down.	An image of a woman in a hat.	The image is a photograph of a woman with a veil.
Idefics-9b Instruct	The image shows a silhouette of a woman with long hair.	A drawing of a woman with long hair and a hat	The image is a portrait of a young girl wearing a bonnet.
LLaVA 1.5 7b	The image is a black and white drawing of a person's face	The image features a woman wearing a black hat and a black coat	The image features a woman wearing a bonnet and a white dress
LLaVA 1.5 13b	The image is a black and white drawing of a woman’s head	The image is a black and white drawing of a woman	The image is a black and white photograph of a woman
mPLUG-Owl	The image features a black and white drawing of a woman with a hat on her head.	The image features a black and white drawing of a woman wearing a hat and a long black dress.	The image features a young girl with long, dark hair wearing a white dress and a white bonnet.
InstructBLIP Flan T5xl	The image features a black and white drawing of a woman with a hat.	a black and white drawing of a woman wearing a fur coat and hat	a black and white drawing of a girl wearing a hat.
BLIP2 OPT2.7b	a woman in a hat	the woman in the hat	the girl in the hat
BLIP2 OPT6.7b	a woman with a hat on her head	-	-
BLIP2 Flan T5xl	a woman with a hat	a woman in a fur coat	a girl with a hat

Figure 23: Young Old woman generative examples



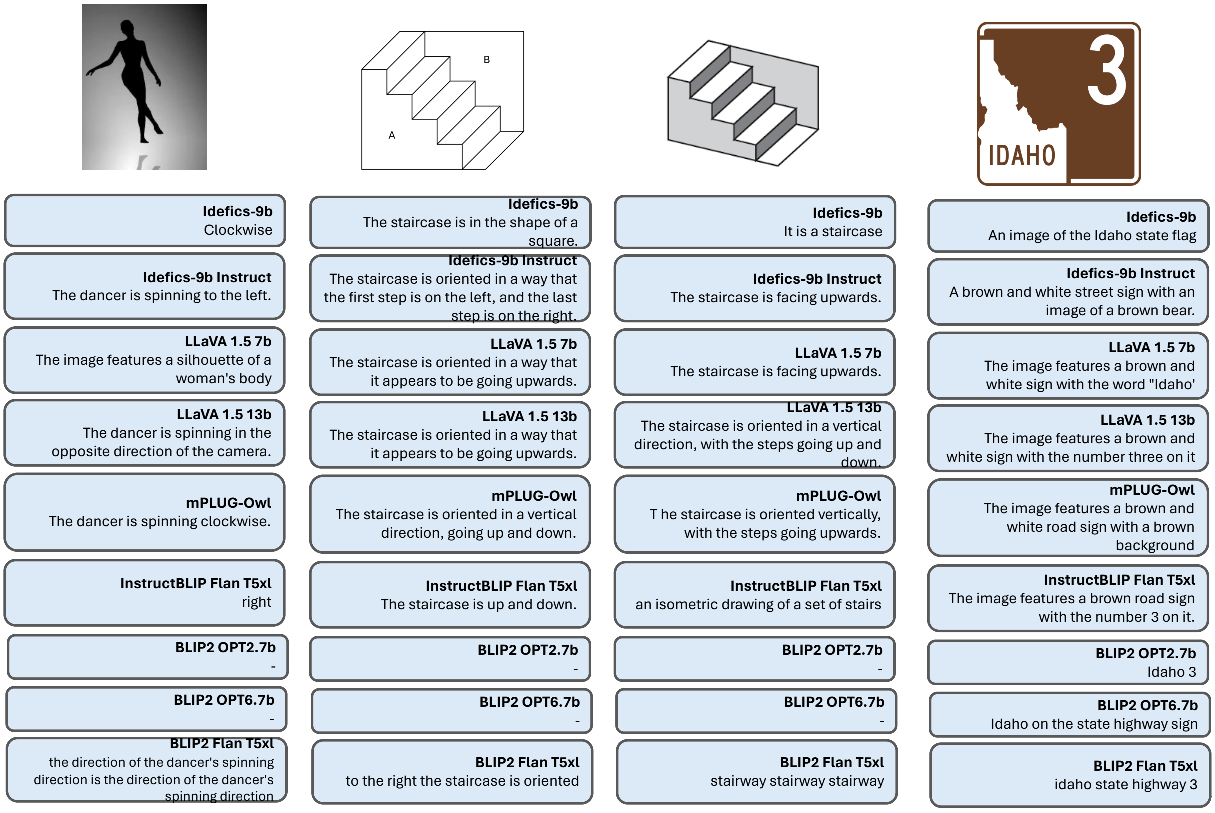
Idefics-9b	An image of a person looking in a mirror.	An image of a man's face.	a wrought-iron lattice tower
Idefics-9b Instruct	The silhouette of a man and a woman's heads together.	a silhouette of a man's head and shoulders	Two faces, one male and one female, are shown on the left side.
LLaVA 1.5 7b	The image features a black and white silhouette of a person	The image features a black and white drawing of a person's face	The image features a black and white drawing of a person's face
LLaVA 1.5 13b	The image features a black and white photo of two men's faces	The image is a black and white drawing of a person's head	The image features a black and white design of two faces,
mPLUG-Owl	The image features a black and white silhouette of two people facing each other.	The image features a black and white drawing of a man's face,	The image features a black and white silhouette of two people's faces.
InstructBLIP Flan T5xl	The image features a silhouette of two men facing each other	The image features a silhouette of a person's head in the shape of a vase.	two silhouettes of heads facing each other on a black background
BLIP2 OPT2.7b	two people in silhouette	-	two people silhouettes
BLIP2 OPT6.7b	two people facing each other	-	two people silhouettes
BLIP2 Flan T5xl	two heads facing each other	a silhouette of a person	-

Figure 24: Vase-Faces woman generative examples



Idefics-9b It is a right-handed cube	**Idefics-9b** It is a right-handed cube	**Idefics-9b** It is a cube
Idefics-9b Instruct The cube is oriented in the shape of a square.	**Idefics-9b Instruct** The cube is oriented in the top-right corner of the image	**Idefics-9b Instruct** The cube is oriented with its top face visible.
LLaVA 1.5 7b The orientation of the cube is such that it is facing upwards.'	**LLaVA 1.5 7b** The orientation of the cube is such that it is facing the viewer.	**LLaVA 1.5 7b** The orientation of the cube is such that it is facing upwards, with the top of the cube visible.
LLaVA 1.5 13b The cube is oriented in a way that it is facing upwards	**LLaVA 1.5 13b** The cube is oriented in a way that it is facing upwards	**LLaVA 1.5 13b** The cube in the image is oriented in a way that it appears to be a square.
mPLUG-Owl The ocube is oriented as a square..	**mPLUG-Owl** The orientation of the cube is square..	**mPLUG-Owl** The cube is oriented horizontally.
InstructBLIP Flan T5xl right	**InstructBLIP Flan T5xl** right side up	**InstructBLIP Flan T5xl** The cube is oriented vertically
BLIP2 OPT2.7b -	**BLIP2 OPT2.7b** -	**BLIP2 OPT2.7b** -
BLIP2 OPT6.7b -	**BLIP2 OPT6.7b** -	**BLIP2 OPT6.7b** -
BLIP2 Flan T5xl the cube is oriented in the direction of the x axis	**BLIP2 Flan T5xl** the cube is oriented in the direction of the x axis	**BLIP2 Flan T5xl** the cube is oriented in the vertical direction

Figure 25: Necker-Cube generative examples on question “What is the orientation of the cube?"



Idefics-9b Clockwise	Idefics-9b The staircase is in the shape of a square.	Idefics-9b It is a staircase	Idefics-9b An image of the Idaho state flag
Idefics-9b Instruct The dancer is spinning to the left.	Idefics-9b Instruct The staircase is oriented in a way that the first step is on the left, and the last step is on the right.	Idefics-9b Instruct The staircase is facing upwards.	Idefics-9b Instruct A brown and white street sign with an image of a brown bear.
LLaVA 1.5 7b The image features a silhouette of a woman's body	LLaVA 1.5 7b The staircase is oriented in a way that it appears to be going upwards.	LLaVA 1.5 7b The staircase is facing upwards.	LLaVA 1.5 7b The image features a brown and white sign with the word "Idaho'
LLaVA 1.5 13b The dancer is spinning in the opposite direction of the camera.	LLaVA 1.5 13b The staircase is oriented in a way that it appears to be going upwards.	LLaVA 1.5 13b The staircase is oriented in a vertical direction, with the steps going up and down.	LLaVA 1.5 13b The image features a brown and white sign with the number three on it
mPLUG-Owl The dancer is spinning clockwise.	mPLUG-Owl The staircase is oriented in a vertical direction, going up and down.	mPLUG-Owl T he staircase is oriented vertically, with the steps going upwards.	mPLUG-Owl The image features a brown and white road sign with a brown background
InstructBLIP Flan T5xl right	InstructBLIP Flan T5xl The staircase is up and down.	InstructBLIP Flan T5xl an isometric drawing of a set of stairs	InstructBLIP Flan T5xl The image features a brown road sign with the number 3 on it.
BLIP2 OPT2.7b -	BLIP2 OPT2.7b -	BLIP2 OPT2.7b -	BLIP2 OPT2.7b Idaho 3
BLIP2 OPT6.7b -	BLIP2 OPT6.7b -	BLIP2 OPT6.7b -	BLIP2 OPT6.7b Idaho on the state highway sign
BLIP2 Flan T5xl the direction of the dancer's spinning direction is the direction of the dancer's spinning direction	BLIP2 Flan T5xl to the right the staircase is oriented	BLIP2 Flan T5xl stairway stairway stairway	BLIP2 Flan T5xl idaho state highway 3

Figure 26: Spinning Dancer results on question “What is the dancer’s spinning direction?", Shroeder Stairs on “What is the orientation of the stairs" and Idaho-Face on "describe the image".