

Visual Interestingness Decoded: How GPT-4o Mirrors Human Interests

Fitim Abdullahu and Helmut Grabner
Zurich University of Applied Sciences, Switzerland
{fitim.abdullahu, helmut.grabner}@zhaw.ch

Abstract

Our daily life is highly influenced by what we consume and see. Attracting and holding one’s attention – the definition of (visual) interestingness – is essential. The rise of Large Multimodal Models (LMMs) trained on large-scale visual and textual data has demonstrated impressive capabilities. We explore these models’ potential to understand to what extent the concepts of visual interestingness are captured and examine the alignment between human assessments and GPT-4o’s, a leading LMM, predictions through comparative analysis. Our studies reveal partial alignment between humans and GPT-4o. It already captures the concept as best compared to state-of-the-art methods. Hence, this allows for the effective labeling of image pairs according to their (commonly) interestingness, which are used as training data to distill the knowledge into a learning-to-rank model. The insights pave the way for a deeper understanding of human interest. Code and materials: <https://github.com/fiabdu/Visual-Interestingness-Decoded>

1. Introduction

Online media data continues to expand, making it increasingly challenging to deliver relevant and engaging content to users. A key aspect of this challenge is the concept of (visual) interestingness – capturing attention and influencing behavior, which dates back to Berlyne’s work in 1949 [2]. On the other hand, a vast amount of online accessible media is scraped to empower the training of foundation models in a self-supervised manner. Large Multimodal Models (LMMs), especially Language-Vision Models like GPT-4o [32], encode human-like knowledge and perform impressively across tasks. While they can reliably categorize images or answer visual questions, their ability to recognize subjective concepts remains uncertain.

This work investigates how well state-of-the-art LMMs capture the fuzzy concept of visual interestingness. We explore whether these models can identify features associated with interestingness and compare their assessments to human judgments through user studies. As illustrated in

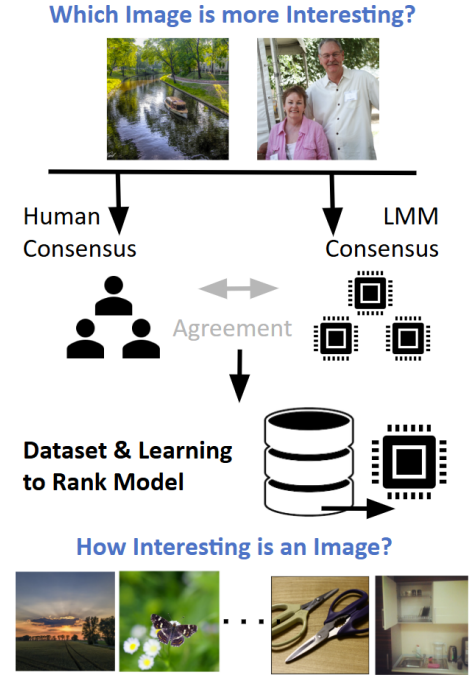


Figure 1. LMMs, such as GPT-4o, encode human-like knowledge and perform well across various tasks. We explore image interestingness, a highly subjective concept, by examining consistent labeling between humans and LMMs and their level of agreement. Pairwise labeling accesses relative measures, which are used to train a rank model to finally assess an image’s interestingness.

Fig. 1, we focus on both (i) the alignment and divergence between human- and model-based evaluations and (ii) assessing LMMs’ potential to reduce manual labeling effort by distilling knowledge from their internal representations.

We focus on everyday images to make the task traceable, consistent with prior work [11, 12, 18]. By achieving high consensus among humans, we explicitly limit subjectivity in the task, aiming to identify images that resonate with a broader audience [1, 5, 16]. We balanced the dataset size with experimental feasibility to ensure traceability and manage computational costs. Despite this trade-off, our results remain significant, and the insights are valid.

Approach	Data	Definition of Interestingness	Labels & Computational Model
IJCV’21 [9] et al.	images, image features and meta-data	explicit human annotations, e.g., AMT	“direct” supervised training using the human annotations
ECCV’24 [1]	images (large scale from Flickr)	user interaction, i.e., user favorites	unsupervised estimation of user agreement to identify commonly interesting images; used to train a regressor
proposed	images (large scale)	LMM trained on a vast amount of (human-generated) data	label pairwise images to capture local preferences, which are then combined into a global learning-to-rank model

Table 1. Most related work has focused on explicit human annotation. However, obtaining reliable annotations at scale is challenging and costly. Recent work explores implicit information to obtain “real” data from photo-sharing platforms at scale. However, this data may be biased toward a particular application, and only positive intentions (i.e., favorites) are available. To overcome these limitations, our approach leverages knowledge represented in LMMs to distill a computational model for visual interestingness.

To this, our main contributions are:

- We present a dataset of 1,000 images (2,500 image pairs) with metadata and multiple verified labels from human annotators and three state-of-the-art LMMs.
- We propose a novel approach to estimating image interestingness without direct user input, which outperforms state-of-the-art models for everyday images.
- We analyze and compare human and LMMs annotations, highlighting areas of agreement and divergence, paving the way for better understanding human and machine interests.

2. Related Work

Interestingness is inherently ambiguous¹. It is subjective, varying by context, observer, and individual background [2, 9, 18]. Yet, recent work suggests that certain image features appeal broadly, regardless of individual differences [1, 5, 16]. Constantin et al. [9] provides a comprehensive overview of computational approaches to visual interestingness. For instance, Flickr calculates an “interestingness” score to help users discover engaging content on its platform [4, 14]. Early research in this area primarily relied on classical machine learning and computer vision methods, often constrained by limited datasets [12, 17, 18]. As the field has grown, the advent of deep learning and more powerful computational resources enabled larger-scale studies and more complex models, e.g., [8].

However, unbiased data collection and consistent annotations at a large scale are challenging and costly. Abdullah and Grabner [1] recently proposed a more progressive approach that explores indirectly labeled data from multiple Flickr users. “Interestingness” is defined by analyzing user engagement, and their approach aims to identify commonly appealing image characteristics directly from users’ favorites. In our work, we aim to extend this idea and get an-

notations from the knowledge encoded into LMMs trained on a vast amount of (human-generated) data – see Tab. 1.

Large Multimodal Models. The emergence of multimodal foundation models has transformed AI by combining computer vision and natural language processing within a unified framework [7, 13, 32, 40, 43]. This paradigm shift changes training approaches and enables broader, more flexible model applications, moving from discrete classification tasks to prompt-based interactions [15]. This flexibility is advantageous in question-answering tasks, where users may ask about unfamiliar or abstract concepts. GPT-4o achieves state-of-the-art performance on various visual perception benchmarks [31, 32].

Recent models like GPT-4o have been trained on vast amounts of web data, encoding extensive knowledge from different fields. Fine-tuning and distilling knowledge into smaller, more specific models is widely used; see [44] for a recent survey. Leveraging the general knowledge of LMMs, specific models are used to design automatic evaluators that mirror human performance. Applications include evaluating text and images (e.g., 3D models) [42], assisting graphic design [19], assessing fashion aesthetics [20], or measuring content appeal [6].

To align with human values, models are fine-tuned with supervision [34] or automatically [36]. However, their latent knowledge can go beyond what LMMs are explicitly taught [25, 28], potentially revealing both overlaps and gaps between human and machine understanding [23, 39]. Understanding these overlaps and gaps could help uncover new concepts and insights², especially in subjective areas.

Outline of the Paper. We aim to explore how the implicit knowledge encoded in LMMs relates to the concept of visual interestingness. Sec. 3 and Sec. 4 present consensus and agreement for single and relative image interestingness assessment, respectively. Sec. 5 uses automatically annotated image pairs from GPT-4o to train a computational model for predicting image interestingness within a learning-to-rank framework. Finally, Sec. 6 discusses the

¹ Ambiguities – i.e., missing information to specify the task explicitly – are common [35]. As a classic example, let’s consider René Magritte’s 1929 painting *The Treachery of Images* [26]. The artwork, depicting an image of a pipe with the caption “Ceci n’est pas une pipe” (“This is not a pipe”), challenges viewers’ perceptions of images and symbols.

² A widely known example might be the classical “Move 37” of AlphaGo when playing against Lee Sedol [29].

similarities and differences between humans and GPT-4o, providing insights into what makes an image interesting.

3. Single Image Interestingness Assessment

3.1. Experimental Design

Dataset. We chose images from the photo-sharing platform Flickr as these images capture a wide variety of content from diverse communities, including professionals and amateur users [1, 24]. To select representative and diverse images, we selected 1,000 images from the Flickr-User dataset [1], equally sampled according to their proposed commonly interestingness score.

Human Annotations. We conducted our user study using Amazon Mechanical Turk. Human workers were instructed on evaluating image interestingness and describing their choice, with no right or wrong answers – it was their decision if an image matched their interest (intentionally kept very open). More specifically, a randomly selected image was shown along with the question “Is this image interesting to you?” with two response options: yes or no. Additionally, they were prompted to provide a brief explanation for their choice. Five Human Intelligence Tasks (HITs) were created for each of the 1,000 images. Each HIT contained one image, and workers were compensated \$0.01 per completed task. A total of 258 unique workers participated, with each worker labeling images without seeing the same image more than once.

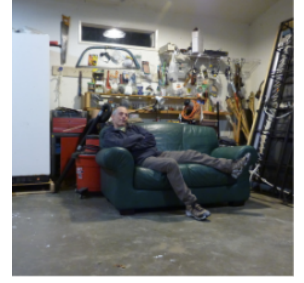
For further analysis, we split the dataset $\mathcal{S} = \mathcal{C}_H \cup \mathcal{D}_H$. An image belongs to $\mathcal{C}_H$ if four or five participants ($(\cdot)_H$, for human) agree on their response, indicating a high level of consensus. Otherwise, it belongs to $\mathcal{D}_H$, reflecting dissent. Additionally, we define the interestingness label y_H as 1 if the majority finds the image interesting and 0 otherwise.

LMM Annotations. The rapid development of foundation models is remarkable, especially considering how quickly new models are released. In our work, we use state-of-the-art LMMs, specifically OpenAI’s GPT-4o [32] (‘gpt-4o-2024-11-20’), Meta’s Llama 3.2 [13] (‘Llama-3.2-11B-Vision-Instruct’), , and DeepSeek’s Vision-Language VL2 [43] (‘DeepSeek-VL2-tiny’). All of which can take text-image inputs and produce text outputs. The models are given the same prompt as the human annotators (the exact prompt is provided at the beginning of the subsequent subsections), with each image being evaluated five times. Analog to human annotations, consistently labeled images belong to sets $\mathcal{C}_G$ (for GPT-4o), $\mathcal{C}_L$ (for Llama 3.2), and $\mathcal{C}_D$ (for DeepSeek) others to the corresponding dissent sets $\mathcal{D}_{\{G,L,D\}}$. Annotation labels $y_{\{G,L,D\}} \in \{0, 1\}$ are defined by majority vote.

Asking for Explanations. Asking “why” is used twofold, (i) to ensure quality for the judgments (see discussion for Llama 3.2 and DeepSeek annotations in Sec. 4.1)



(a) **Human:** 5/5 yes; Explanations: “ILIKE”, “It is good photo and interesting”, “It is amazing”, ...
GPT-4o: 5/5 yes; Explanations: “The vibrant sunset and scenic landscape create a captivating visual appeal.”, “The sunset and landscape create a visually stunning scene.”, “The sunset with radiant clouds over a vast field creates a captivating and serene scene.”, ...



(b) **Human:** 5/5 yes; Explanations: “NEAT”, “Thinking”, ...
GPT-4o: 4/5 yes; Explanations: “The cluttered workshop with a person lounging hints at a story.”, “It depicts a unique juxtaposition of a living space and a workshop.”, “The image depicts a person relaxing in a cluttered garage, which might not appeal to everyone.”, ...

Figure 2. Examples of images and corresponding responses from human annotators and GPT-4o. Almost all images are consistently labeled, most of them as interesting.

	$ \mathcal{C}_x $	$y_x = 1$	$A^{(H,x)}(\mathcal{S})$	$A^{(H,x)}(\mathcal{C}_H)$
Human	91.9 %	99.9 %	-	-
GPT-4o	93.9 %	95.3 %	92.9 %	93.6 %
Llama	93.1 %	99.8 %	97.1 %	98.3 %
DeepSeek	76.2 %	81.4 %	75.3 %	77.3 %

Table 2. Consistency and agreement with human annotation for single image interestingness assessment. Unfortunately almost all images are consistently labeled as interesting.

and (ii) to gain deeper insights about how humans and LMMs come to the particular conclusion whether the image is interesting (see Sec. 6). Examples are shown in Fig. 2.

3.2. “Is This Image Interesting?”

Prompt: *Is this image interesting? Answer with one word (yes or no) without punctuation and in lowercase. Add a semicolon without space. Explain why in one sentence without going into detail.*

As summarized in Tab. 2, the annotations from humans and LMMs show high consistency (almost all images are in the respective consistent set $\mathcal{C}_x$). This indicates that humans and LMMs generally agree on the interestingness of images. Furthermore, almost all images in the respective sets are considered interesting ($y_x = 1$), indicating that humans and LMMs find almost all images on which they agree to be interesting. These suggest that humans and LMMs actively look for something interesting when explicitly asked for it, leading to a predominantly positive response.

The agreement $A^{(M,N)}(\mathcal{S}) := \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \mathbb{I}(y_M^{(i)} = y_N^{(i)})$ between annotation M and N on set $\mathcal{S}$ measures how well annotations of humans and LMMs are aligned. Not surprisingly, as almost all images are interesting, the results indicate a high level of agreement between humans and LMMs. Furthermore, the agreement between human and LMM increases slightly when focusing on consistently labeled images in $\mathcal{C}_H$. It seems that it is somewhat easier for the LMM to distinguish between interesting and uninteresting if the humans agree on this question.

Notably, the DeepSeek model has significantly less consistency (among itself) and less agreement with humans. However, for humans and the other LMMs, some images fall into the dissenting sets. This inconsistency may arise because an image may not interest a broad audience. For example, in Fig. 2b, GPT-4o stated: “The image depicts a person relaxing in a crowded garage, *which might not appeal to everyone.*” This response suggests that the model subjectively evaluates the content to determine its interestingness, as discussed in [1].

Key Insight. Responses from humans, GPT-4o, and Llama 3.2 are very consistent and aligned. Almost all images were deemed interesting, and a story was made up to support the decision.

4. Relative Image Interestingness Assessment

As demonstrated in the last sections, whether an image is interesting is hard to answer generally. Results on single images are rendered meaningless, as almost all images are consistently labeled as interesting by humans and state-of-the-art LMMs. Relative comparisons are more affordable and often used for similar judgments, e.g., [18, 41].

4.1. Experimental Design

Dataset. We created image pairs based on the 1,000 images used previously. Each image was used in five different (random) pairs, resulting in 2,500 image pairs.

Human Annotations. As in the previous experiment, Amazon Mechanical Turk was used to obtain human annotations. A randomly selected image pair was shown to a worker, who was asked: “Which image is more interesting to you?” The worker could click on their preferred image and was asked to briefly explain their choice. Consistent with the previous experiment, five HITs were created for each of the 2,500 image pairs. Each HIT contained one image pair, and workers were compensated \$0.01 per completed task. A total of 553 unique workers participated, with each worker labeling image pairs without seeing the same pair more than once.

As above, the consistency of the answers is defined if four or more humans agree on the labeling ($\mathcal{C}_H$). Furthermore, let y_H represent the human labels, where $y_H = 1$



Human: 5/5 second; Explanations: “Love Ferraris!”, “Very nice”, “It looks good”, ...

GPT-4o: 5/5 second; Explanations: “The vibrant color and modern design stand out more.”, “The modern, sleek design of the vehicle coupled with the vibrant color captures attention more effectively.”, “It’s visually striking due to its modern design and vivid color.”, ...



Human: 5/5 second; Explanations: “I like birds”, “This bird is cute”, “LOOKING NATURAL”, ...

GPT-4o: 5/5 first; Explanations: “The intricate design and craftsmanship make it more visually engaging.”, “It showcases a unique and artistic representation.”, “The first image depicts a unique and intricate metallic insect sculpture, making it more visually striking.”, ...

Figure 3. Image pairs illustrating instances where humans and GPT-4o agree and disagree. For example, humans and GPT-4o have differing opinions regarding images of insects and birds. At first glance, it may not be immediately evident that the insect image is a metallic sculpture, which could explain why people did not find it as interesting—humans may not give the image the same level of attention as a machine.

if the majority prefers the first image and $y_H = 0$ if the majority prefers the second image.

GPT-4o Annotations. In this study, two image inputs are used, which is supported by GPT-4o.

Systematic Error. Even though the model allows multiple images as input, we have discovered a systematic error. Image pairs were presented twice to GPT-4o, and the images were swapped on the second run. For 36% of the cases, GPT-4o always reported the second image as more interesting. Only 64% (1,599 out of 2,500) of the GPT-4o annotation remained the same, independent of the image order. For subsequent experiments, only these image pairs were kept. Please note that this systematic error does not seem to be much correlated to the human consensus (56.3%

of image pairs in $\mathcal{C}_H$ and 47.6% in $\mathcal{D}_H$ are error-free).

GPT-4o Demographics. As interestingness is subjective, it would be nice to test different user groups based on their demographics automatically. We used the system prompt of GPT-4o for that purpose: “*You are a [gender] from [continent] and between [age] and [age] years old.*” If one uses prominent images that have become naturalized in society for men or women, such as a blue car or a pink flower, GPT-4o responds differently. A car is more interesting for men and a pink flower for women, somehow capturing prejudices. However, this vanishes when using everyday images where this distinction is no longer so prominent. In a more extensive study involving 500 random image pairs, we used *male* or *female* for gender and *North America* or *Africa* for the continent, and a range of 25 to 34 and 45 to 54 years for age, respectively. Running all eight combinations, filtering out pairs with a systematic error, and combining the remaining pairs, we ended up with 116 image pairs. Unfortunately, the results were identical for all image pairs, regardless of gender, continent, or age specified.

Llama 3.2 and DeepSeek-VL2 Annotations. As Llama 3.2 [13] currently does not support multiple image inputs, we combined the two images into a single input for the model. However, this workaround did not yield reliable results. For instance, the model often selected the first image as more interesting while providing an explanation that referred to the second image. Similarly, DeepSeek’s recent Janus Pro Model [7] does not support multiple image inputs. When combining images, the model consistently selected the second image as more interesting. While DeepSeek’s Vision-Language V2 model do allow for multiple image inputs [43], the selection and descriptions often do not align with the actual content, exhibiting issues similar to those observed with Llama 3.2.

Due to these inconsistencies, which result in an unfair comparison between these LMMs and GPT-4o, we limited our further analyses to GPT-4o.

Key Insight: GPT-4o responds independently of the demographic tested; however, it has a significant systematic bias in favoring the second image over the first.

4.2. “Which image is More interesting?”

Prompt: *Which of the two images is more interesting? Answer with one word (first or second) without punctuation and in lowercase. Add a semicolon without space. Explain in one sentence why you have chosen this image without going into detail.*

Unlike in the single-image study above, people’s responses are less consistent in the paired-image experiment. Set $\mathcal{C}_H$ contains 56.3% of the image pairs, indicating consensus in about half of them. GPT-4o exhibits much higher consistency, with 95.5% of all image pairs in $\mathcal{C}_G$. The overall agreement between GPT-4o and human annotations is

$A^{(H,G)}(\mathcal{S}) = 66.2\%$. In case of human consensus, the agreement increases to $A^{(H,G)}(\mathcal{C}_H) = 73.8\%$, while decreases on the dissent set to $A^{(H,G)}(\mathcal{D}_H) = 56.5\%$ – as one would expect. Examples are depicted in Fig. 3.

Key Insight. GPT-4o’s annotations are aligned with human judgments, especially when there is consensus among people.

4.3. Comparisons and Relation to other Approaches

Our annotations are compared to other models and concepts related to visual interestingness prediction (c.f. [9]). All approaches provide a measurement, score, or probability per image, which (is claimed) to be related to the image’s interestingness. An image pair is labeled according to which image yields the higher response.

Aesthetics [22]. The VILA (Vision-Language Aesthetics) model learns image aesthetics by analyzing user comments alongside images. It models subjective aesthetic judgments by aligning visual features with language and categorizing images according to these learned aesthetics.

Memorability [21]. A predictive model assigns memorability scores to images. Through large-scale experiments with memory recall tasks, quantified memorability is established as a stable metric across viewers. We use the recent re-implementation from [30].

Commonly Interesting Images (CI) [1]. Interestingness is subjective. However, some images appeal to a broader audience and are, therefore, of common interest. A predictive model was trained by analyzing how many unique Flickr users “favored” images from a certain category (i.e., visually similar images).

Social Interestingness [11]. Whereas being related to visual interestingness, factors beyond image features are relevant to make an image go viral. Social interestingness metrics use the number of views, favorites, and comments of a post on a social media platform. Using Flickr images in our study, we directly sourced these values for every image.

Zero shot learning [37]. We use Customized Prompts via Language (CuPL) to generate prompts to determine an image’s interestingness. Using a Large Language Model, in our case GPT-4o, to “Describe what an *interesting* image looks like”, we get various prompts such as “An interesting image features vibrant colors, unexpected elements, and a captivating composition that draws the viewer’s eye”. Overall, 500 prompts are created following [37]. After removing duplicates and highly similar prompts, we ended up with 250 unique prompts. Text embeddings for these prompts using CLIP [38] are calculated and averaged. The final score is the cosine similarity between text and image CLIP embeddings.

Results and Discussion. Results can be seen in Tab. 3 (left). Every approach is compared to human (y_H) and GPT-4o annotations (y_G), in terms of agreement on the set

Group	Model	Annotations (Sec. 4)		Learning to Rank (Sec. 5)			
		$A^{(H,x)}$	$A^{(G,x)}$	$Acc.^{(H)}$	$r_S^{(H)}$	$Acc.^{(G)}$	$r_S^{(G)}$
Human	Human	-	73.8 %	$77.5 \pm 2.5 \%$	-	$72.0 \pm 3.4 \%$	0.59 ± 0.06
LMMs	GPT-4o	73.8 %	-	$73.4 \pm 3.4 \%$	0.59 ± 0.06	$84.8 \pm 2.5 \%$	-
	CuPL	60.3 %	60.9 %	$61.5 \pm 3.5 \%$	0.34 ± 0.07	$63.2 \pm 3.1 \%$	0.42 ± 0.08
Computational Models	CI	69.6 %	67.6 %	$69.6 \pm 3.6 \%$	0.54 ± 0.06	$69.1 \pm 3.3 \%$	0.52 ± 0.06
	Memorability	35.5 %	39.1 %	$34.7 \pm 4.0 \%$	-0.42 ± 0.08	$38.3 \pm 3.6 \%$	-0.34 ± 0.07
	Aesthetic	68.3 %	75.1 %	$69.0 \pm 3.7 \%$	0.50 ± 0.07	$73.6 \pm 3.7 \%$	0.67 ± 0.06
Social Inter- estingness	#Views	61.7 %	63.9 %	$63.4 \pm 3.4 \%$	0.39 ± 0.08	$66.3 \pm 3.4 \%$	0.48 ± 0.08
	#Favorites	66.4 %	74.0 %	$66.3 \pm 3.2 \%$	0.47 ± 0.07	$69.4 \pm 3.1 \%$	0.57 ± 0.07
	#Comments	68.0 %	74.8 %	$66.6 \pm 3.1 \%$	0.46 ± 0.07	$70.2 \pm 3.2 \%$	0.58 ± 0.07

Table 3. GPT-4o achieves the highest agreement $A^{(\cdot,x)}$ among all models using human responses as the ground truth. It also outperforms existing models focused on visual interestingness, related concepts, and social metrics. On the right, we show the model’s accuracy $Acc.^{(\cdot)}$ on the image pairs. The global ranking (measured by the Spearman rank correlation $r_S^{(\cdot)}$) of the test dataset remains consistent, indicating that the learning-to-rank model generalizes beyond pairwise relationships for both human and GPT-4o annotations.

$\mathcal{C}_H$. GPT-4o is superior to previous models in this context, followed by models using aesthetic or common interestingness. Social interestingness scores reveal that images with more comments tend to be considered more interesting than those without, which aligns with research in that regard [10]. Memorability score has the weakest link to interestingness, consistent with prior findings [18]. When GPT-4o is used as the ground truth, the VILA aesthetic model performs the best, followed by the GPT-4o model, while the human model ranks third. It also appears to have a stronger agreement with social interestingness, possibly due to their pre-training.

Key Insight. GPT-4o’s annotations are superior to previously proposed approaches to predict human interest.

5. Learning a Computational Model

So far, image pairs have been annotated by humans, GPT-4o, and various computational models. In this section, we distill this knowledge into a simple computational model.

Learning-To-Rank. A simple learning-to-rank model can be implemented using a Siamese network architecture with shared weights [3]. As we are using images $\mathbf{I}_0$ and $\mathbf{I}_1$ as input, they are first embedded using CLIP³, passing through a linear layer with a single neuron. The scoring function is the difference between them, passed through a sigmoid function: i.e., $S(\mathbf{I}_0, \mathbf{I}_1) := \sigma(\mathbf{w}^\top \text{CLIP}(\mathbf{I}_0) - \mathbf{w}^\top \text{CLIP}(\mathbf{I}_1))$. Learning is done to maximize the score differences between pairs. Binary cross-entropy loss on the

target $y \in \{0, 1\}$ is used; $y = 1$ if $\mathbf{I}_0$ ranks higher than $\mathbf{I}_1$ and $y = 0$ otherwise. As the weights are shared, after training, a score can also be obtained using a single input $S(\mathbf{I}) = \sigma(\mathbf{w}^\top \text{CLIP}(\mathbf{I}))$. For multiple images, the individual scores are used to rank them. Besides its simplicity, this approach has been used successfully many times, also for distilling information from LLM or LMM, e.g., [6, 41]

Training/ Testing. The dataset was split in half for training and testing. We train learning-to-rank models for all annotations (human and GPT-4o and approaches from Sec. 4.3). Each model was trained for 25 epochs, and no overfitting was observed. Each experiment was repeated 50 times with different training/ test splits. The results are depicted in Tab. 3 (right). $Acc.$ denotes the model accuracy on the individual image pairs (as trained) and r_S the Spearman rank correlation on the global ranking based on the scores $S(\cdot)$ for each image.

Results and Discussion. The best-performing model is obtained when training and test data are from the same source (human or GPT-4o). This serves as the baseline for comparing the other models. All the results match nicely with those from the previous sections (left side of the table) for the individual performance of labeling image pairs.

According to human annotations, the model generalizes to unseen data, although the average accuracy of approximately 77.5% may reflect the subjectivity of the task. GPT-4o achieves the best performance among the models, although a gap remains compared to the baseline. Other computational models, such as CI or aesthetics, perform well but still fall short of GPT-4o’s results. Examining the Spearman correlations, we find that human responses positively correlate with all models except memorability, which aligns with current research findings. Notably, the correlation between humans and GPT-4o is 0.59, indicating a moderate positive

³We also perform experiments with DINOv2 [33] embeddings. Similar, slightly worse (67.1% for Humans and 68.3% for GPT-4o) results and trends were achieved. This might be because CLIP was trained on text-image pairs, which provided some supervision, whereas DINOv2 is trained purely in a self-supervised manner on images. For more details, see the supplementary material.

relationship between these two variables.

Based on GPT-4o annotations, the aesthetic model ranks highest in accuracy after GPT-4o itself, followed closely by the human model. The global ranking of the test dataset is inline, meaning that the learning-to-rank model can generalize beyond pairwise relationships for both human and GPT-4o annotations.

Int10k Dataset. The Int10k dataset [9] focuses on video summarization, and there is a significant domain gap between *single everyday images* (our focus in this work) and cinematic image sequences. Nevertheless, we applied our approach. Overall, the results show a significant drop, with accuracy for the human model and human-provided annotations being around $59.3\% \pm 2.5\%$. Even for this dataset, GPT-4o outperformed all other approaches, achieving a comparable accuracy of $59.2\% \pm 2.1\%$. For more details, see the supplementary material.

Key Insight. Computational models obtained from distilling the information from the annotations match them very well. GPT-4o is superior to previous models.

6. Similarities and Differences in Assessment

This section discusses the agreement and disagreement between humans and GPT-4o in more detail.

Embeddings. We analyzed the responses from both GPT-4o and human annotators to create a semantic embedding using OpenAI’s ‘text-embedding-3-small’ model (1536-dimensional vector for each text input). Since the human responses were predominantly short and repetitive – often consisting of simple terms like “nice”, “beautiful”, or “good” – they were less suitable for in-depth analysis (cf. Fig. 4). Therefore, we focused our analysis on the text embeddings from GPT-4o responses.

Exploiting the “Why” responses. We perform hierarchical clustering on the embeddings. Several clusters emerged from the analysis, including the *cute* and *emotional* clusters shown in Fig. 4a, which are of interest to both humans and GPT-4o. While humans respond to certain images as *cute*, GPT-4o tends to associate them with emotions. Additionally, humans and GPT-4o demonstrate consensus regarding *uniqueness*. Nevertheless, there are clusters that only GPT-4o finds interesting, such as images featuring *vibrant color* or depicting action. These images are not necessarily captivating for humans, see Fig. 4b. Please note that these clusters represent semantically similar texts, not semantically similar images.

Exploiting Image Appearance. To gain deeper insight into what makes an image interesting or uninteresting, we conducted an additional experiment focusing on the visual characteristics of the images. GPT-4o was asked to describe each image, and embeddings were calculated for those descriptions. Fig. 5a-5b illustrates a hierarchical clustering of the description embeddings in the original space and a 2d



Human: “Cute Girl.”, “This side is cute”, “THIS IMAGE IS VERY BEAUTIFULL”, “It looks interesting”

GPT-4o: “It captures a joyful moment.”, “It shows human interaction and activity.”, “The image of the baby animals captures a unique and lively moment.”, “The emotional connection between the two beings adds an engaging element.”



Human: “Right is better than left”, “This building’s design is very interesting.”, “Appreciate nature more”, “What a two wheeler excited to ride immediately”

GPT-4o: “Features unique graffiti on a monument.”, “The building’s unique architecture makes it stand out.”, “The image depicts a unique and intriguing structure situated in a picturesque and seemingly remote location.”, “The vintage motorcycle has a unique and nostalgic appeal.”

(a) Clusters considered as “interesting” for humans and GPT-4o.



GPT-4o: “The action and dynamic scene with soldiers makes it more interesting.”, “It has more vibrant colors and dynamic elements”, “The arrangement and colors are visually appealing.”, “The image shows a person performing music on the street which adds more dynamic and action.”

(b) Only “interesting” for GPT-4o (no human descriptions are available, as these images were never considered interesting in pairwise comparisons by human annotators).

Figure 4. Explanations: The clusters are derived from text embeddings of responses “Why” an image is interesting. Clusters of common interest (a) include “cute”, “joyful” moments as well as “uniqueness”. A minority of clusters (b) shows disagreement between humans and GPT-4o.

projection using UMAP [27]. The clustering reveals groups of semantically similar images consistently found interesting, uninteresting, or divisive by humans and GPT-4o, see Fig. 5c-5e. There is agreement on images depicting flowers, birds, or nature scenes as interesting. In contrast, humans and GPT-4o find images of mundane scenes or people in ordinary situations uninteresting. Images depicting people at events or performing art elicit mixed responses.

Tab. 4 reports quantitative information per cluster, including agreement between humans and GPT-4o and if the

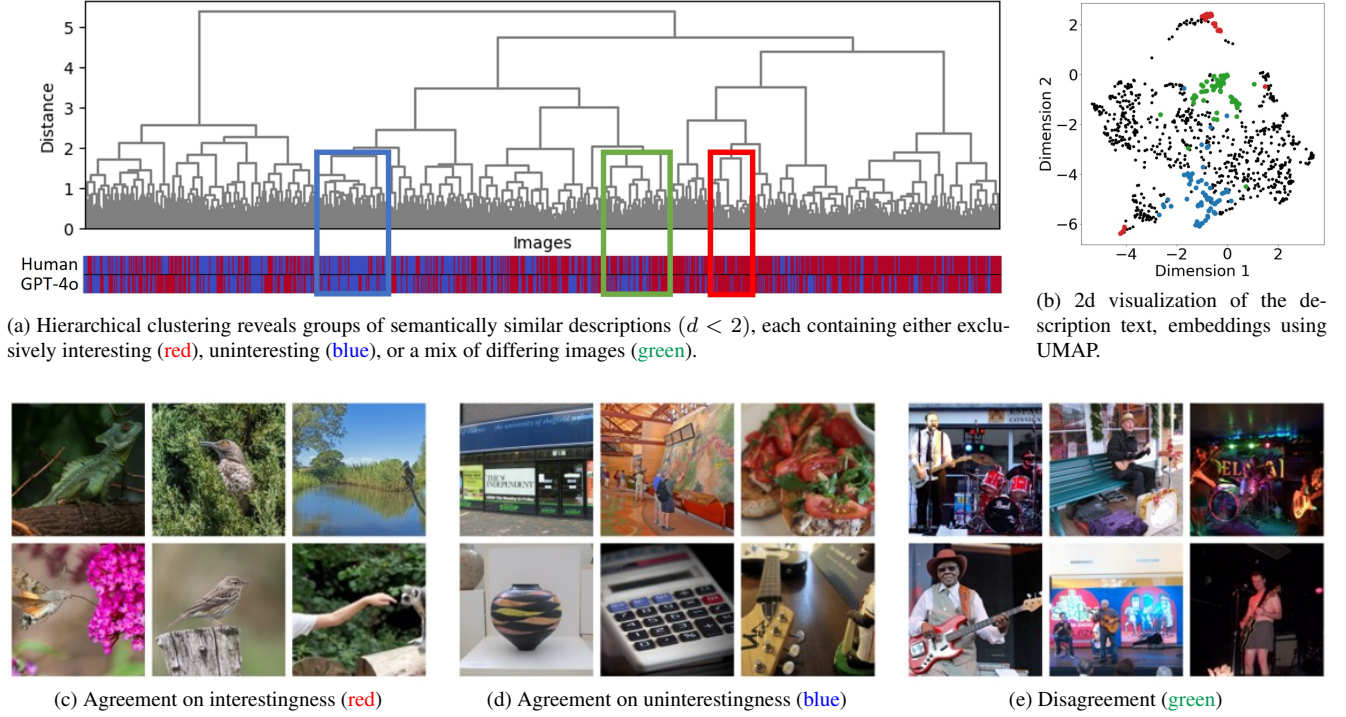


Figure 5. Image Content Descriptions: Text embeddings of the image descriptions are divided into semantically similar groups using hierarchical clustering (a, b). Most clusters indicate agreement between humans and GPT-4o (c, d), while some indicate disagreement (e). Compare with Tab. 4.

cluster is mainly interesting or uninteresting. Additionally, we calculated the mean ranks of images in each cluster based on both human ($\bar{R}^{(H)}$) and GPT-4o ($\bar{R}^{(G)}$) annotations. The mean ranks align closely, particularly when agreement is high. The top four words describe the cluster obtained after removing stop words and lemmatizing the automatically generated appearance descriptions. These words align well with examples from Fig. 5 (marked).

Key Insight. Humans and GPT-4o generally agree well on what is interesting for topics and scenes.

7. Conclusion

GPT-4o cannot be used directly to assess image interestingness due to uninformative (almost always positive) responses on single images and systematic errors. However, we demonstrated that GPT-4o outperforms existing models in a comparative annotation setting. Its alignment with human assessments highlights its potential to support (i) large-scale studies and (ii) knowledge distillation. Further investigation into the gap between GPT-4o and human annotation, including demographic factors, would be insightful. Extending the study to datasets beyond Flickr images could provide a broader perspective. Together, these efforts will help refine our understanding of visual interestingness and its contributing factors.

A	$\frac{\#A_{pos}}{\#A}$	$\bar{R}^{(H)}$	$\bar{R}^{(G)}$	Frequent Words (Appearance)
94%	60%	143	184	train, track, station, railway
86%	23%	246	236	people, group, front, standing
84%	86%	86	107	water, sky, background, body
82%	97%	53	80	perched, bird, branch, flower
81%	24%	217	207	person, sitting, room, window
81%	59%	165	142	building, modern, street, large
80%	46%	204	193	people, two, smiling, together
80%	78%	125	131	dog, lying, person, cat
78%	88%	72	79	tree, sky, landscape, water
77%	18%	223	221	various, small, featuring, ...
77%	41%	202	191	person, wearing, standing, red
72%	44%	212	158	person, people, background, ...
72%	89%	74	91	water, bird, swimming, white
71%	61%	197	158	people, group, person, flag
74%	54%	154	185	car, parked, red, background
65%	72%	95	160	flower, green, yellow, pink
47%	71%	207	163	playing, stage, performing, guitar
40%	50%	207	163	stage, guitar, person, group

Table 4. Agreement of humans and GPT-4o concerning the image content; agreement (A) on **un-** and **interestingness**, average ranks from models trained using human and GPT-4o annotations as well as **disagreement** along its clusters text descriptions. Color markings match Fig. 5.

Acknowledgements. This research was funded by the Swiss National Science Foundation (SNSF) under grant number 206319 “Visual Interestingness – All images are equal but some images are more equal than others”.

- [1] Fitim Abdullahu and Helmut Grabner. Commonly interesting images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 4, 5
- [2] Daniel E Berlyne. Interest as a psychological concept. *British Journal of Psychology*, 39(4):184, 1949. 1, 2
- [3] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the International Conference on Machine learning (ICML)*, pages 89–96, 2005. 6
- [4] Daniel S Butterfield, Catherine Fake, Callum J Henderson-Begg, and Serguei Mourachov. Interestingness ranking of media objects, USPTO #US8732175B2. 2
- [5] Fabienne Bünzli, Wibke Weber, Fitim Abdullahu, and Helmut Grabner. Depicting humans, animals, and objects in motion: The effect of implied motion on engagement and persuasion in advertising. *Journal of Advertising*, 0(0):1–21, 2024. 1, 2
- [6] Sherry X. Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Misha Sra, and Pradeep Sen. Aid-appeal: Automatic image dataset and algorithm for content appeal enhancement and assessment labeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 6
- [7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2, 5
- [8] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. Computational understanding of visual interestingness beyond semantics: Literature survey and analysis of covariates. In *ACM Computing Surveys*, volume 52. Association for Computing Machinery, 5 2019. 2
- [9] Mihai Gabriel Constantin, Liviu Daniel Ștefan, Bogdan Ionescu, Ngoc Q.K. Duong, Claire H  l  ne Demarty, and Mats Sj  berg. Visual Interestingness Prediction: A Benchmark Framework and Literature Review. *International Journal of Computer Vision (IJCV)*, 129(5):1526–1550, 5 2021. 2, 5, 7
- [10] Lisette de Vries, Sonja Gensler, and Peter S.H. Leeﬂang. Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing*, 26(2):83–91, 2012. 6
- [11] Arturo Deza and Devi Parikh. Understanding image virality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1818–1826, 2015. 1, 5
- [12] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages
- Spataru, Baptiste Koziere, Bethany Biron, Binn Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnston, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yearry, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur C  lebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiao-

qing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelen, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg

Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojuan Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 3, 5

- [14] Flickr. About interestingness, <https://www.flickr.com/explore/interesting/>, 2024. 2024-02-23. 2
- [15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26. Curran Associates, Inc., 2013. 2
- [16] Maham Gardezi, King Hei Fung, Usman Mirza Baig, Mariam Ismail, Oren Kadosh, Yoram S Bonneh, and Bhavin R Sheth. What makes an image interesting and how can we explain it. *Frontiers in Psychology*, 12, 2021. 1, 2
- [17] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. Visual interestingness in image sequences. In *Proceedings of the ACM International Conference on Multimedia*, pages 1017–1026, 2013. 2
- [18] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1633–1640, 2013. 1, 2, 4, 6
- [19] Daichi Haraguchi, Naoto Inoue, Wataru Shimoda, Hayato Mitani, Seiichi Uchida, and Kota Yamaguchi. Can gpts evaluate graphic design based on design principles? In *SIG-*

- GRAPH Asia 2024 Technical Communications*, 2024. 2
- [20] Yuki Hirakawa, Takashi Wada, Kazuya Morishita, Ryotaro Shimizu, Takuya Furusawa, Sai Htaung Kham, and Yuki Saito. An empirical analysis of gpt-4v’s performance on fashion aesthetic evaluation. In *SIGGRAPH Asia 2024 Technical Communications*, 2024. 2
- [21] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011. 5
- [22] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5
- [23] Been Kim. Beyond interpretability: developing a language to shape our relationships with ai, Apr 2022. 2
- [24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 2020. 3
- [25] Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 34651–34663. Curran Associates, Inc., 2022. 2
- [26] René Magritte. The treachery of images (this is not a pipe) (la trahison des images [ceci n’est pas une pipe]), 1929. 2024-10-24. 2
- [27] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7
- [28] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17359–17372. Curran Associates, Inc., 2022. 2
- [29] Cade Metz. In two moves, alphago and lee sedol redefined the future. *WIRED.com*, 16, 2016. 2
- [30] Coen D Needell and Wilma A Bainbridge. Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*, 5(2):168–184, 2022. 5
- [31] OpenAI. Hello gpt-4o, <https://openai.com/index/hello-gpt-4o/>, 2024. 2024-07-01. 2
- [32] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Rei-ichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sasstry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Val-

- lone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024. 1, 2, 3
- [33] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6
- [34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. 2
- [35] Alicia Parrish, Susan Hao, Sarah Laszlo, and Lora Aroyo. Is a picture of a bird a bird? a mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 1–18, Torino, Italia, May 2024. ELRA and ICCL. 2
- [36] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, abs/2406.16855, 2024. 2
- [37] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 5
- [39] Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero, 2023. 2
- [40] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gura, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Meray, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hod-

kinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturk, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjöstrand, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Borral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White,

Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohanane, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklín, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinta Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dekhtyarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bølle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon,

Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjit Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellen-doorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jijing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaime Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irv-

ing, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumathi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHosseini Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mo-

jtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzasczcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauer, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Ram-mohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshv, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesch Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listfk, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Za-

her, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhan, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Dannenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohmman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2024. 2

- [41] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison. *arXiv preprint arXiv:2402.16641*, 2024. 4, 6
- [42] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [43] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 2, 3, 5
- [44] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024. 2