

Modeling Collaborator: Enabling Subjective Vision Classification With Minimal Human Effort via LLM Tool-Use

Imad Eddine Toubal^{1,2*} Aditya Avinash¹ Neil Gordon Alldrin¹ Jan Dlabal¹ Wenlei Zhou¹
 Enming Luo¹ Otilia Stretcu¹ Hao Xiong¹ Chun-Ta Lu¹ Howard Zhou¹
 Ranjay Krishna^{1,3}† Ariel Fuxman¹ Tom Duerig¹

¹Google Research ²University of Missouri ³University of Washington

itdfh@umsystem.edu, {adity, nalldrin}@google.com

Abstract

From content moderation to wildlife conservation, the number of applications that require models to recognize nuanced or subjective visual concepts is growing. Traditionally, developing classifiers for such concepts requires substantial manual effort measured in hours, days, or even months to identify and annotate data needed for training. Even with recently proposed Agile Modeling techniques, which enable rapid bootstrapping of image classifiers, users are still required to spend 30 minutes or more of monotonous, repetitive data labeling just to train a single classifier. Drawing on Fiske’s Cognitive Miser theory, we propose a new framework that alleviates manual effort by replacing human labeling with natural language interactions, reducing the total effort required to define a concept by an order of magnitude: from labeling 2,000 images to only 100 plus some natural language interactions. Our framework leverages recent advances in foundation models, both large language models and vision-language models, to carve out the concept space through conversation and by automatically labeling training data points. Most importantly, our framework eliminates the need for crowd-sourced annotations. Moreover, our framework ultimately produces lightweight classification models that are deployable in cost-sensitive scenarios. Across 15 subjective concepts and across 2 public image classification datasets, our trained models outperform traditional Agile Modeling as well as state-of-the-art zero-shot classification models like ALIGN, CLIP, CuPL, and large visual question answering models like PaLI-X.

1. Introduction

The field of computer vision has primarily focused on recognizing concepts that are objectively agreed upon, such

*This work was done during an internship at Google.

†This work was done during working at Google.

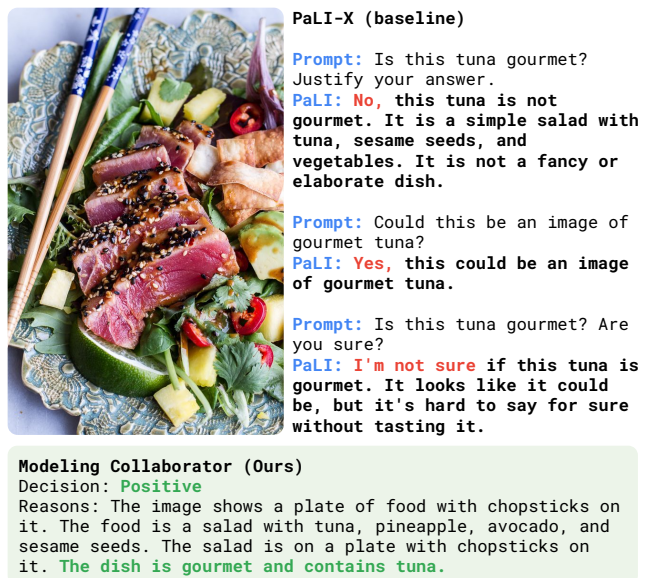


Figure 1. We introduce Modeling Collaborator: a framework that allows anyone to train vision models using natural language interactions and minimal effort. We show that today’s best models (e.g. PaLI-X [6]) change their answers depending on the prompt when classifying subjective concepts like gourmet tuna. Meanwhile, Modeling Collaborator uses LLMs and tool-use to train vision models by interacting with users to carve out the concept space.

as dogs, cats, or cars [11, 28, 30]. Even research on fine-grained recognition (e.g. “black footed albatross”) and compositional concepts (e.g., “red car next to a motorcycle”) have universal consensus [22, 27, 32, 35]. However, many practical real-world vision applications frequently involve recognizing subjective concepts that suffer from significant disagreements amongst individuals. Applications include predicting emotions, measuring aesthetic appeal, or content moderation [10, 25, 26, 45]. A content moderator

needs a model to identify unsafe content according to their definition of what constitutes as *unsafe*; a food critic might not consider a tuna sandwich to be gourmet while others might (Figure 1). To operationalize these applications, we need user-centric training frameworks that enable anyone to train subjective vision models.

Recently, Agile Modeling formalized the process for turning any visual concept into a vision model through a user-in-the-loop framework [51]. Their work concluded that crowd workers struggled to produce labels that were consistent with the user’s concept definition. Instead, they proposed an active learning algorithm, where the user iteratively labels a series of training images themselves. Unfortunately, this process is tedious, repetitive, and labor intensive; users had to label ~ 2000 images, which on average took 30 minutes to train a binary classifier.

Existing processes fall short because they do not leverage a key capability that humans possess. People are adept at breaking down complex subjective concepts into more manageable and objective components by applying first-order logic [14, 36]. This ability can be explained using Susan Fiske’s Cognitive Miser Theory: people decompose complex work to avoid high cognitive load [13]. People apply the same process to define complex concepts such as “unsafe” and “gourmet”. For instance, one food critic might decompose the subjective concept of “gourmet” as images that need to at least contain “tuna”; if it is “ahi tuna”, then it is likely gourmet; if it is “canned”, then it is unlikely to be gourmet; if the dish is a “sandwich”, then it is still not gourmet. This decomposition of the subject concept “gourmet” into conjunction clauses of objective concepts “ahi tuna”, “canned”, and “sandwich” is a simple *non-laborious*, cognitively effortless conversion.

With this grounding, we deliver **Modeling Collaborator** which empowers users to build classifiers while minimizing manual effort. Instead of asking users to annotate thousands of images [51], Modeling Collaborator requires 100, along with a few natural language interactions that help decompose subjective concepts into its constituent sub-components. To enable Modeling Collaborator, we leverage advancements in large language models (LLMs) [2, 3, 9, 12, 37] and in particular, their ability to use vision-language models (VLMs) [6–8] and other tools [19]. When users have a concept in mind and use Modeling Collaborator, it employs an LLM, which breaks the concept into questions that are digestible for a Visual Question Answering (VQA) model [8]. The LLM then summarizes the answers provided by the VQA model and performs reasoning through chain-of-thought [57] to classify new images as positive or negative examples of the concept. Users are only asked to manually label a small 100 image validation set. Finally, Modeling Collaborator labels a large amount of unlabeled images available online and uses it as distillation

data to train a light-weight deployment-ready vision model.

Our method is shown to outperform existing zero-shot methods (CLIP [43], CuPL [41] and PaLI-X [6]), especially on harder subjective concepts. When compared to the original Agile Modeling [51] our system exceeds the quality of crowd-raters on hard concepts while simultaneously reducing the need for manual user-provided ground-truth by orders of magnitude. By reducing the barriers of manual effort and resulting costs needed to develop classification models, it will empower users to rapidly convert their ideas into reality. This, in turn, has the potential to usher in a new wave of end-user applications.

2. Related work

Our work draws on advances in VLMs and LLMs and provides an improved solution to the recently introduced Agile Modeling problem.

Agile Modeling. Inspired by agile software development, Agile Modeling [51] focuses on rapid development of image classification models. In addition to speed, Agile Modeling aims to tackle the challenges posed by subjective vision models. As classification tasks become more nuanced, user interaction becomes increasingly crucial. However, it is important to note that the human-in-the-loop approach can be expensive due to the need of continuous human involvement and expertise. While this work aims at reducing time users spend on tuning their classification models, we propose an assisted method to automate parts of the pipeline and eliminate crowd-rater involvement.

Vision-language models (VLMs). In the rapidly evolving domain of VLMs, two primary streams have emerged: contrastive and generative models. Contrastive models, such as CLIP [43] and ALIGN [23], leverage large-scale datasets to directly learn visual concepts from raw text, enabling high-accuracy zero-shot classification on open vocabularies [11, 17]. Generative models such as PaLI [6–8, 56] and GPT-V [37, 38] focus on generating text from a combination of visual and text inputs. For instance, PaLI, trained on a vast collection of image-text pairs in various languages, achieves top performance across a range of vision and language tasks. Similarly, GPT-V allows the processing of image inputs, thereby enhancing the applicability of language models to multimodal tasks. Other methods such as CoCa [54, 63] proposed a hybrid approach for simultaneously learning with generative and contrastive objectives. Despite their strength, VLMs capture visual data semantics, often prioritizing salient image features over nuanced visual cues. For instance, CLIP embeddings are intentionally compressed to encapsulate its most prominent subject [49]. Additionally, PaLI may struggle to provide detailed descriptions of complex scenes with numerous objects, as its training data predominantly lacks detailed annotations. In contrast, our proposed method is more stable and less sensitive

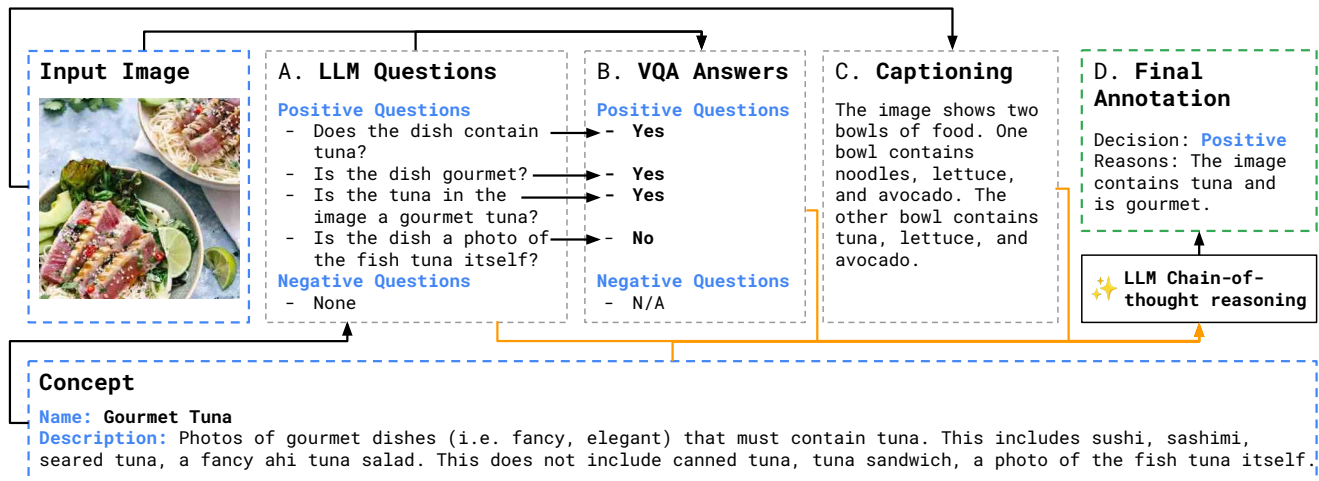


Figure 2. Modeling Collaborator Annotator system. For a given image, concept name, and description, the Annotator outputs a positive or negative label. Based on the name and description of the concept, the LLM generates relevant atomic questions to ask a VQA model (PaLI VQA in our case) (step A). These questions are fed into the VQA model that typically outputs a yes/no short answer (Step B). Additionally, we use a captioning version of PaLI (Step C) to generate a detailed description capturing as much detail as possible from the image. Finally, the LLM goes through a chain-of-thought reasoning process to output a decision and rationale (Step D).

to question phrasing as observed in Fig. 1.

Large language models (LLMs) and tool-use. Large Language Models (LLMs) have revolutionized the landscape of artificial intelligence [1, 3, 12, 40, 55], particularly in the field of natural language processing (NLP) and cognitive reasoning. By leveraging advanced methodologies such as chain-of-thought reasoning [57], few-shot learning [4, 39], and tool-use [21, 46], these models demonstrate exceptional performance across a wide spectrum of downstream tasks [44]. They can operate across various modalities and a broad range of applications while maintaining high performance without the need for additional training. Recent progress in integrating external tools with LLMs [5, 20, 21, 29, 62] has yielded systems like Toolformer [46]. This approach makes intelligent decisions about which APIs to invoke, optimizing the timing, arguments passed, and the subsequent assimilation of the results into future token predictions. This enhances zero-shot performance across a variety of tasks, establishing a solid foundation for LLMs to operate beyond their inherent capabilities. For fine-grained VQA, AVIS [20] introduces an autonomous information-seeking mechanism. By dynamically leveraging an LLM in tandem with external tools, it adeptly traverses a combinatorial search space. This is achieved through its unique approach of mimicking human decision-making processes, crafting a transition graph that guides the LLM’s strategic decisions. Another tool-use enabled LLM system is ViperGPT [52], which embodies an innovative approach to tackling visual queries. It leverages a code-generation strategy that enables the seamless integration of vision-and-language models through the generation of Python code. This method,

along with other similar methods (MMReact [61], HuggingGPT [50], Chameleon [34], and Visual ChatGPT [58]) circumvents the need for extended training and ensures resilience across a diverse set of visual tasks. Collectively, these systems highlight the burgeoning synergy between LLMs and external tool use, pushing the frontiers of what LLMs can achieve. In our work, we adopt and extend ideas from these approaches to tackle subjective classification.

Customized prompts via language models. Customized Prompts via Language models (CuPL) [41] leverages CLIP’s capabilities [43] to achieve zero-shot image classification. CuPL measures the similarity between an image and each visual class to perform classification. Typically, the classes are passed into CLIP’s text encoder within a template such as “photo of a bird” for the class bird. CuPL employs GPT [3] to generate more comprehensive text descriptions for each class before feeding into CLIP. This straightforward and zero-shot approach yields improved accuracy across various zero-shot image classification benchmarks. However, its evaluation has been limited to objective classification tasks and not on nuanced or subjective visual classification tasks. This approach for automatically annotating data improves upon CLIP but suffers from the same limitations compared to our work.

3. Method

We propose an end-to-end system that streamlines the development of classifiers for nuanced visual concepts, addressing the limitations of traditional classifier development methods. The system consists of three core components, de-

scribed in detail in the following subsections: (a) data mining, (b) annotation, (c) model training with active learning.

To build a classifier for a new concept, the user first provides a concept name and an optional description. The system then automatically mines images relevant to the concept and annotates them using a mixture of Large Language Models (LLM), Vision-Language Models (VLM), and Visual-Question-Answering (VQA) models. The annotated images are used to train a basis classification model, which is further refined through multiple rounds of active learning, resulting in a highly accurate classifier.

This setup mirrors the workflow of traditional classifier development, but it eliminates the need for costly and time-consuming human annotation which is a significant bottleneck in traditional methods. The Modeling Collaborator Annotator component, powered by LLMs and VLMs, enables zero-shot image labeling and drastically minimizes our dependence on user annotations.

3.1. Data mining

Mining quality data for training has traditionally been a labor-intensive process. This process begins with the clear definition of a concept, followed by the hunt for relevant images, and ends in the manual annotation of each of these images [11, 30]. Particularly for nuanced visual tasks, there is a possibility that certain subtle visual patterns might be overlooked during data collection. Consequently, to ensure a comprehensive capture of all visual patterns, multiple iterations of refinement may be needed. In traditional Agile Modeling [51] this challenge is addressed by soliciting *users* to annotate data or generate new search queries to find more image examples. Each query results in a new semantic image search algorithm [23, 43] to gather other similar positive image examples for annotation from the public domain (LAION Dataset) [47]. Even with user intervention, *user* queries may overlook essential cues, potentially leading to a deficit of hard negatives or a lack of coverage in specific visual modes. Additionally, the labels can vary between users, leading to potential human biases.

To address human bias and minimize manual effort, we propose a data mining algorithm based on LLM chain-of-thought reasoning. While LLMs are not inherently unbiased [15] and may reflect biases present in their training data, they can assess a wider range of concepts at large scales from their extensive knowledge base, thus identifying a broader array of potential examples more efficiently. First, we prompt the LLM to generate multiple positive and negative queries based on a concept’s name and its description. Note that we do not directly assign images as positive or negative based on the query; rather, the goal is obtain representative images spanning both positive and hard-negative examples. To increase coverage and diversity, we expand the queries by instructing the LLM to apply various *mu-*

tations. For example, we may ask the LLM to iteratively come up with broader or narrower versions of the queries, or come up with variations for specific parts of the queries. Drawing parallels to Agile Modeling, we use each query to extract image samples from the public domain [47].

3.2. Modeling Collaborator Annotator

Fig. 2 describes the image annotation process. Our system effectively orchestrates the annotation process leveraging LLM’s ability to invoke VLMs and other tools. It comprises three primary AI-driven modules: an LLM, a Captioning VLM [56], and a VQA VLM [6]. The automated annotation process is structured as follows:

Concept initialization: Initially, our system receives a concept name (e.g., `gourmet tuna`), and optionally a concept description. If a concept description is absent, the LLM generates an initial description. This template can be modified by the user to cover all specifications and carve-outs.

Attribute Extraction: Based on the concept specifications, the LLM identifies objective attributes associated with the concept, such as “image contains tuna”, “is tuna sandwich”, and “is tuna steak”.

Attribute decomposition: The LLM decomposes complex attributes into more granular and atomic attributes.

Question generation: The LLM then formulates a series of questions tailored for the VQA model. Examples include “does the image contain food”, “is the food tuna”, and “is it tuna steak”.

Visual assessment: When an image is input, the VQA model processes these questions, yielding concise answers for each. Concurrently, the Captioning VLM provides a comprehensive description of the image.

Final annotation: With the textual data from the VLMs and the user’s initial concept specification, the LLM employs chain-of-thought reasoning. It annotates the image as either positive or negative, also offering insights into its decision-making process.

Our approach utilizes the strengths of VLM, VQA, and LLM models while simultaneously avoiding their shortcomings. For example, VLMs, despite their capabilities, often struggle with nuanced and subjective concepts in classification tasks. Their performance hinges on the breadth and quality of training data, potentially leading to biases or gaps in understanding [53]. Ambiguities in language and the inherent subjectivity of certain questions can further challenge their accuracy [33]. Moreover, these models, lacking real-world context and experiential understanding, might miss deeper cultural or emotional nuances [16]. Thus, while powerful, VLMs have inherent limitations in addressing intricate or subjective visual-linguistic tasks. Fig. 1 shows an example VLMs’ (PaLI-X [6]) sensitivity to prompts.

VLMs are primarily designed for understanding and answering questions related to visual content, rather than

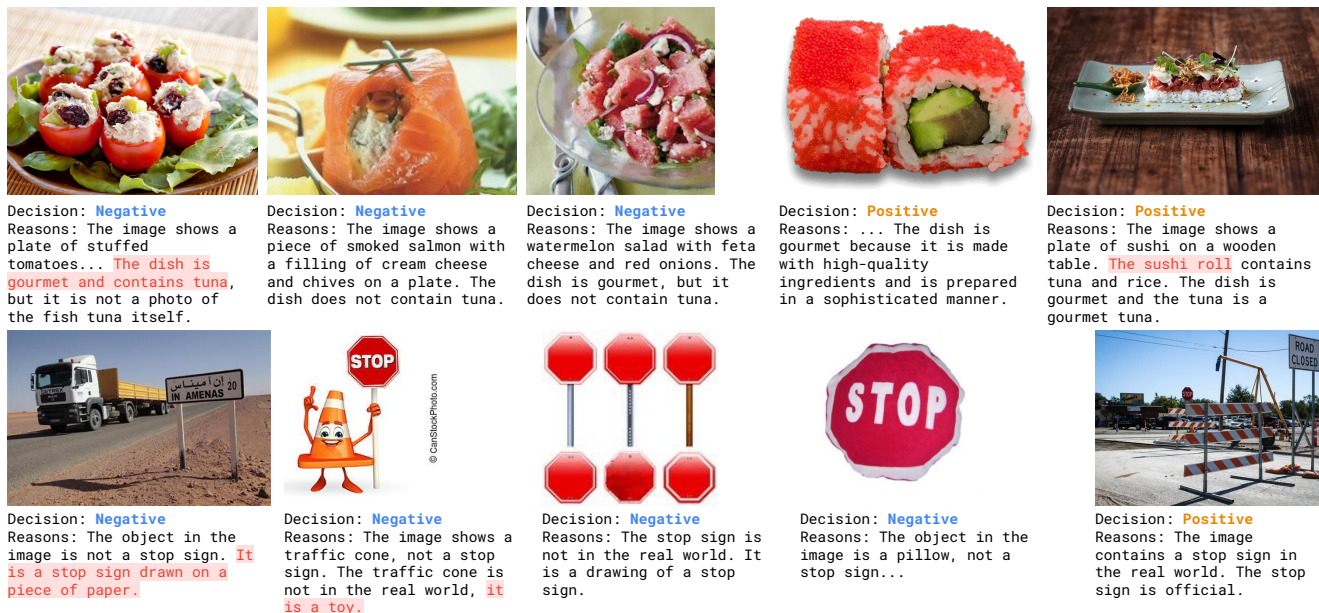


Figure 3. Modeling Collaborator Annotator examples for the concepts `gourmet tuna` (first row) and `stop sign` (second row). Hard negatives mined from the LAION dataset are shown in addition to some actual positives for the visual concept. The Modeling Collaborator Annotator is able to label the images as positive or negative as well as provide rationale. In some instances, the rationale could be incorrect (highlighted in red) due to error in VQA responses or hallucinations from the LLMs. Some of the reasons have been truncated for brevity.

performing deep chain-of-thought reasoning typical of advanced LLMs [33,42,53,59]. While VLMs can comprehend simpler questions about images, they usually operate in a single-shot manner, providing answers based on the immediate visual and textual inputs without extended reasoning. On the other hand, LLM question answering quality can be significantly improved through chain-of-thought reasoning, maintaining a coherent line of thought across extended text. Other techniques such as prompt chaining involve using a model’s output as part of the subsequent input, simulating a sustained dialogue or iterative reasoning. Additionally, to extract deeper insights, users can guide LLMs with specific instructions, such as asking the model to think step-by-step [60] or weigh pros and cons, thus simulating a more deliberate reasoning process [3].

3.3. Training and active learning

While one could directly use the annotator as a model, this is prohibitive in many scenarios because of the high inference cost. For this reason, we adopt an approach similar to [51] for model training and active learning. Specifically, we first extract image features from a foundation vision model (CLIP or ALIGN) [23, 24]. We then train a shallow multi-layer perceptron (MLP) with layer sizes (128, 128, 128) to perform binary classification for the given concept. This can also be viewed as student-teacher distillation [18] where we use the LLM-based annotator as the teacher model. We use a learning rate of 3×10^{-4} , a

batch size of 512, and optimize using AdamW [31].

After the initial model is trained, we perform multiple rounds of active learning. Each active-learning iteration consists of three stages. First, the lightweight classification model is applied to a large database of unlabeled images (LAION [47]). Then, we perform stratified sampling to acquire candidate images for further AL rounds [51]. The intention is to capture hard negatives and hard positives that will boost precision and recall respectively. Second, our LLM-based annotator is autonomously applied to the selected images, providing additional training ground-truth. Thirdly, the student classifier is retrained, leveraging all the extant labeled data. We experiment with both margin sampling and stratified sampling techniques [48] to mine examples during this active learning phase. The overall system thus adeptly balances between exploration (achieved via data mining through text search queries and expansion) and exploitation (achieved via active learning to mine visual modes that reduce model uncertainties).

3.4. Implementation details

As a large language model, we use PaLM 2 [2, 9] which was trained on a variety of different tasks, all of which helps PaLM 2 learn different aspects of language. Additionally, we use both the VQA and MMIT (multimodal instruction-tuned [56]) variants of PaLI-X [6]. The particular choice of foundation models is based on their SOTA performance at the time of writing. These models have not been further

trained or fine-tuned in this work.

4. Experiments

We present our experimental setup and results with three takeaways. First, we show that Modeling Collaborator Annotator outperforms other zero-shot methods (CLIP [43], CuPL [41] and PaLI-X [6]). Second, while Modeling Collaborator Annotator is able to beat state-of-the-art methods in both easy and hard concepts, we see much larger gains on harder and more subjective concepts. Finally, when using our end to end system, we can produce deployable models of competitive quality with minimal *user* annotations (100 annotations vs. 2,000 in traditional Agile Modeling).

Datasets. In addition to the LAION dataset used for data mining in our system, we evaluate our methods on the public Hateful Memes dataset [26]. For evaluation and user-study, we use the Agile Modeling dataset [51] that is comprised of 14 concepts, each with positive and negative images mined from the LAION dataset. This dataset is split into *easy* and *hard* concepts depending on the zero-shot performance on each concept using CLIP as described in [51].

Models. We benchmark Modeling Collaborator Annotator against state-of-the-art zero-shot and open-vocabulary classifiers: CLIP [43], CuPL [41], and PaLI-X (55B) [6] as a generative VQA model. We evaluate CLIP by embedding the name of the concept and measuring the cosine similarity to each image embedding. CuPL uses the same technique but instead of embedding the concept name directly, we embed a description of the concept generated by an LLM. Both GPT3 and PaLM 2 models were experimented with but we chose PaLM 2 since it produced superior results. In the case of CLIP and CuPL, we select an operating point using a grid search maximizing the F1 score on a subset of the training set. We use PaLI-X VQA variant as a classifier by prompting it “*Is this an image of X?*” and we assign a positive or negative prediction based on its answer.

Annotator Adaptation. While testing the system, we observed some amount of concept-dependent variability in the Annotator. For example, for simple concepts like “cat” a VLM might already have state-of-the-art performance and our system can even degrade quality in these cases. To address this we implemented six different Annotator *strategies*. While developing a classifier for a particular concept, we have the concept owner build an on-the-fly validation set of 100 images which is then used to select the best performing strategy for that particular concept. Different parameters describing these configurations are explained in the Supplementary Materials.

Users, Crowd, and Modeling Collaborator. We measure the agreement/alignment with the *user* for both the crowd and automatic annotation methods. The *user* is the source of ground-truth and the person manually annotating the test set. *Crowd* annotators are given a description and exam-

ples by the *user* and asked to annotate images at a larger scale. *Modeling Collaborator Annotator* is able to scale up the annotation process further due to its autonomy and can encapsulate an image set of higher diversity in visual modes. We measure the annotator alignment by comparing the performance (auPR) on the distilled model trained on data annotated by different human and machine annotators.

4.1. Modeling Collaborator Annotator

Modeling Collaborator Annotator outperforms other zero-shot methods. We show the results of these experiments in Tab. 1. We measure the alignment with the *user* on the held-out test set of the Agile Modeling dataset using agreement scores (precision, recall, and F1). CLIP and CuPL contrastive models suffer from very low precision in favor of high recall. PaLI-X outperforms contrastive models, making it more suitable as a baseline for our proposed Annotator. **We achieve significant gains for subjective (hard) concepts while maintaining equivalent performance for less subjective (easy) concepts.** Tab. 1 shows a significant skew in concept improvement: over 25% of the concepts showed an F1 score gain of 4% or higher, including hateful memes [26] at 15%, healthy-dish at 6%, and stop-sign at 5%, exhibiting substantial improvements in areas requiring more subjective classifications. This trend indicates that our model is particularly effective for complex or subjective concepts, but may offer only marginal benefits for concepts that PaLI-X is already good at. Regardless, a Wilcoxon Signed-Rank Test on the F1 scores comparing our system against PaLI-X yields a statistically significant improvement across all concepts ($p < 0.01$). In addition to classification, our system outputs rationales shown in Fig. 3.

4.2. Human-machine alignment

Modeling Collaborator can produce deployable models of competitive quality with minimal user annotations. We measure the effect of using varying levels of human and automated annotation in Tab. 2. We note that, while our model cannot exceed the distilled user model performance (distilled on 100% accurate annotations), we can outperform crowd-raters. Our Annotator system significantly outperforms crowd-raters on harder more nuanced concepts (different of 6%). Whereas it slightly under-performs on easy concepts. This is likely due to prediction errors from automated VQA models (PaLI-X) where humans show better performance. In comparison to using other state-of-the-art open-vocabulary zero-shot annotators (CLIP, CuPL and PaLI-X), our system outperforms these methods on both easy and hard concepts. Our fully automated system successfully generates distilled models that match the quality of ones crafted with classical Agile Modeling, with performance within a 2% margin of the *user*’s output. Fig. 4

Concept	PaLI-X [6]			CLIP [43]			CuPL [41]			Ours		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Easy concepts												
arts-and-crafts	0.71	0.97	0.82	0.68	0.86	0.76	0.68	0.90	0.77	0.96	0.75	0.84
dance	0.57	0.87	0.69	0.51	0.95	0.66	0.52	0.89	0.66	0.67	0.95	0.79
emergency-service	0.67	0.88	0.76	0.53	0.87	0.65	0.54	0.91	0.67	0.88	0.73	0.76
hair-coloring	0.76	0.97	0.85	0.70	0.99	0.82	0.70	0.99	0.82	0.76	0.97	0.85
in-ear-headphones	0.70	0.96	0.81	0.43	0.95	0.59	0.44	0.96	0.60	0.82	0.86	0.82
pie-chart	0.80	0.96	0.88	0.52	0.80	0.63	0.50	0.92	0.65	0.80	0.96	0.88
single-sneaker	0.65	0.92	0.76	0.51	0.99	0.67	0.51	1.00	0.67	0.70	0.88	0.78
Easy concepts average	<u>0.69</u>	<u>0.93</u>	<u>0.80</u>	0.55	0.92	0.68	0.56	0.94	0.69	0.80	0.87	0.82
Δ										+11%	-6%	+2%
Hard concepts												
astronaut	0.61	0.87	0.71	0.40	0.95	0.56	0.42	0.95	0.58	0.72	0.79	0.72
block-tower	0.45	0.97	0.62	0.38	0.99	0.55	0.37	0.98	0.54	0.89	0.68	0.66
gourmet-tuna	0.52	0.95	0.67	0.29	1.00	0.45	0.29	1.00	0.45	0.52	0.95	0.67
hand-pointing	0.56	0.99	0.71	0.39	0.87	0.54	0.39	0.94	0.55	0.89	0.79	0.74
healthy-dish	0.38	1.00	0.55	0.37	0.99	0.54	0.38	1.00	0.55	0.84	0.61	0.61
home-fragrance	0.57	0.51	0.54	0.40	0.95	0.56	0.40	0.96	0.57	0.57	0.51	0.54
stop-sign	0.61	0.99	0.76	0.48	1.00	0.65	0.49	0.99	0.65	0.83	0.83	0.81
Hard concepts average	<u>0.53</u>	<u>0.90</u>	<u>0.65</u>	0.39	0.96	0.55	0.39	0.97	0.56	0.75	0.74	0.68
Δ										+22%	-16%	+3%
Overall average	<u>0.61</u>	<u>0.92</u>	<u>0.72</u>	0.47	0.94	0.62	0.47	0.96	0.62	0.78	0.79	0.74
Δ										+17%	-13%	+2%
Hateful memes [26]	<u>0.66</u>	<u>0.42</u>	<u>0.51</u>	0.49	0.98	0.66	0.50	0.87	0.64	0.58	0.77	0.66
Δ										-8%	+35%	+15%

Table 1. Teacher performance (Precision, Recall, and F1 scores). Modeling Collaborator outperforms state-of-the-art zero-shot methods including CLIP, CuPL, and visual query answering models (PaLI-X). Underlined results represent the baseline (PaLI-X) with which our performance is compared to (deltas). We bold the best precision, recall, and F1 for easy concepts, hard concepts and Hateful memes dataset.

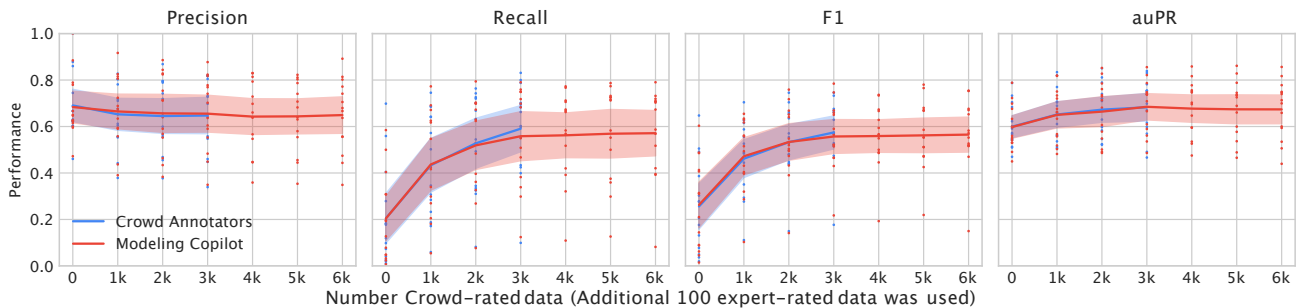


Figure 4. Comparing the contribution of increasingly more training examples annotated by crowd-annotators vs. Modeling Collaborator Annotator (fully automated). The y-axis shows the performance of the final distilled model. When *user* feedback is minimal (100 annotated examples), more crowd-annotators examples improve the final distilled model despite the noisy prediction. Modeling Collaborator Annotator provides similar improvement of performance without any human interactions and can be scaled better to annotate a lot more examples due to its autonomy.

shows that both crowd-annotators and Modeling Collaborator Annotator can improve the performance of the distilled model, even when *user* feedback is minimal. However, Modeling Collaborator Annotator has the advantage of being fully automated and can scale to a larger number

of examples.

Modeling Collaborator and other zero-shot and classical methods fail in complex visual tasks that require complex understanding and reasoning. The effectiveness of our method on identifying hateful memes [26], as demon-

Concept	Human Annotators			Machine Annotators		
	User	Crowd	Crowd	CuPL	PaLI-X	Ours
Dataset size (per concept)	~600	~600	~3000	~3000	~3000	~3000
Easy concepts						
arts-and-crafts	0.77	0.73	0.86	0.78	0.77	0.78
dance	0.69	0.70	0.81	0.72	0.68	0.68
emergency-service	0.75	0.71	0.78	0.59	0.66	0.72
hair-coloring	0.85	0.85	0.83	0.77	0.58	0.80
in-ear-headphones	0.73	0.66	0.67	0.65	0.73	0.72
pie-chart	0.77	0.76	0.76	0.72	0.82	0.82
single-sneaker	0.74	0.64	0.68	0.51	0.61	0.56
Easy concepts average	0.76	<u>0.72</u>	0.77	0.68	0.69	0.73 (+1%)
Hard concepts						
astronaut	0.67	0.71	0.66	0.60	0.65	0.65
block-tower	0.59	0.58	0.45	0.48	0.49	0.50
gourmet-tuna	0.50	0.51	0.35	0.54	0.52	0.52
hand-pointing	0.50	0.56	0.58	0.56	0.81	0.81
healthy-dish	0.59	0.49	0.47	0.42	0.45	0.53
home-fragrance	0.62	0.60	0.69	0.56	0.53	0.53
stop-sign	0.70	0.57	0.55	0.62	0.51	0.64
Hard concepts average	0.60	<u>0.57</u>	0.54	0.54	0.57	0.60 (+3%)
Overall average	0.68	<u>0.65</u>	0.65	0.61	0.63	0.66 (+1%)

Table 2. Quality comparison of different annotators (or teacher models) using the final distilled model performance (auPR). Concept owners provide the highest quality annotations because of their deep understanding of the nuanced concept. Modeling Collaborator annotator provides better quality labels compared with labor-intensive annotations from crowd raters, and compared to other automated methods.

Method	Labeler	# Ex.	F1	Acc	Pre	Rec
Ours (Teacher)	-	-	0.66	0.61	0.58	0.77
CLIP [43]	-	-	0.57	0.53	0.51	0.65
CuPL [41]	-	-	0.51	0.64	0.50	0.87
PaLI-X [6]	-	-	0.51	0.61	0.66	0.42
Ours (Student)	MC	7K	0.56	0.52	0.50	0.64
CLIP+MLP	Human	8.5K	0.48	0.60	0.65	0.38

Table 3. Performance of our method (both Annotator and distilled models) on the Hateful Memes [26] public dataset. Zero-shot and VQA methods are used for comparison.

strated in Tab. 3, is further highlighted by its ability to match fully-trained models without relying on labeled data. Both the teacher and student models outperform the traditional training approach without using any of the training datasets. However, the performance is still low, demonstrating the limitations of our approach.

5. Limitations

As our system is an orchestration of LLMs and VLMs, it can suffer from some of the limitations of its atomic components (PaLM 2, PaLI-X, and CLIP). For example, we observed that providing verbose and overly-complex descriptions of simple concepts (cats, dogs, etc.) can actually de-

grade performance in comparison to simply using PaLI-X. Another issue is that for certain concepts, the CLIP features can lead to poor distilled model quality. One example is `stop sign` (where the stop sign is expected to be a real stop sign in traffic), where the CLIP feature could capture the overall semantics of stop signs, but could not easily discriminate between physical instances vs depictions.

6. Conclusion

In this paper, we presented Modeling Collaborator, a novel framework that alleviates the manual effort required to develop classifiers for subjective and nuanced visual concepts. Our framework leverages advancements in large language models (LLMs) and vision-language models (VLMs) to carve out the concept space through conversation and by automatically labeling training data points. We demonstrate the effectiveness of our framework through a set of experiments, showing that it can quickly build visual classifiers for nuanced concepts and outperform both traditional Agile Modeling and state-of-the-art zero-shot classification models. Our work has the potential to significantly reduce the time and effort required to develop classifiers for a wide range of applications including content moderation and aesthetic classification.

References

- [1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023. [3](#)
- [2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. [2](#), [5](#)
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. [2](#), [3](#), [5](#)
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. [3](#)
- [5] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making with multimodal large language model: Explorations with gpt4-vision and beyond. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023. [3](#)
- [6] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [7] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. [2](#)
- [8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2022. [2](#)
- [9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. [2](#), [5](#)
- [10] Tee Connie, Mundher Al-Shabi, and Michael Goh. Smart content recognition from images using a mixture of convolutional neural networks. In *IT Convergence and Security 2017: Volume 1*, pages 11–18. Springer, 2017. [1](#)
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. [1](#), [2](#), [4](#)
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. [2](#), [3](#)
- [13] Susan T Fiske and Shelley E Taylor. *Social cognition*. McGraw-Hill Book Company, 1991. [2](#)
- [14] Gottlob Frege et al. Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. *From Frege to Gödel: A source book in mathematical logic*, 1931:1–82, 1879. [2](#)
- [15] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023. [4](#)
- [16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. [4](#)
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [2](#)
- [18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. [5](#)
- [19] Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Tool documentation enables zero-shot tool-usage with large language models. *arXiv preprint arXiv:2308.00675*, 2023. [2](#)
- [20] Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David A Ross, Cordelia Schmid, and Alireza Fathi. Avis: Autonomous visual information seeking with large language models. *arXiv preprint arXiv:2306.08129*, 2023. [3](#)
- [21] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *CVPR*, pages 23369–23379, 2023. [3](#)
- [22] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020. [1](#)
- [23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. [2](#), [4](#), [5](#)
- [24] Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. Graph-rise: Graph-regularized image semantic embedding. *arXiv preprint arXiv:1902.10814*, 2019. [5](#)
- [25] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. *ICCV*, pages 2390–2398, 2015. [1](#)

- [26] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 33:2611–2624, 2020. **1, 6, 7, 8**
- [27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. **1**
- [28] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4. *ICCV*, 128:1956–1981, 2020. **1**
- [29] Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. Lmeyer: An interactive perception network for large language models. *arXiv preprint arXiv:2305.03701*, 2023. **3**
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. **1, 4**
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. **5**
- [32] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016. **1**
- [33] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *NeurIPS*, 29, 2016. **4, 5**
- [34] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *NeurIPS*, 36, 2024. **3**
- [35] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *CVPR*, pages 10910–10921, 2023. **1**
- [36] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956. **2**
- [37] OpenAI. Gpt-4 technical report, 2023. **2**
- [38] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. Accessed: 2023-11-15. **2**
- [39] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022. **3**
- [40] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. **3**
- [41] 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*, pages 15691–15701, 2023. **2, 3, 6, 7, 8**
- [42] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022. **5**
- [43] 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 *ICML*, pages 8748–8763, 2021. **2, 3, 4, 6, 7, 8**
- [44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1:9, 2019. **3**
- [45] Arpita Roy, Anamika Paul, Hamed Pirsiavash, and Shimeji Pan. Automated detection of substance use-related social media posts based on image and text analysis. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 772–779. IEEE, 2017. **1**
- [46] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. **3**
- [47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. **4, 5**
- [48] Burr Settles. Active learning literature survey. 2009. **5**
- [49] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *ICLR*, 2021. **2**
- [50] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *NeurIPS*, 36, 2024. **3**
- [51] Otilia Stretcu, Edward Vendrow, Kenji Hata, Krishnamurthy Viswanathan, Vittorio Ferrari, Sasan Tavakkol, Wenlei Zhou, Aditya Avinash, Enming Luo, Neil Gordon Alldrin, et al. Agile modeling: Image classification with domain experts in the loop. *ICCV*, 2023. **2, 4, 5, 6**
- [52] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. **3**
- [53] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. **4, 5**
- [54] Imad Eddine Toubal, Yi-Ting Chen, Krishnamurthy Viswanathan, Daniel Salz, Ye Xia, and Zhongli Ding. Multi-

- modal dual-tower architectures for entity retrieval from image and text. In *CVPRW*, 2023. 2
- [55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3
- [56] Yaqing Wang, Jialin Wu, Tanmaya Dabral, Jiageng Zhang, Geoff Brown, Chun-Ta Lu, Frederick Liu, Yi Liang, Bo Pang, Michael Bendersky, et al. Non-intrusive adaptation: Input-centric parameter-efficient fine-tuning for versatile multimodal modeling. *arXiv preprint arXiv:2310.12100*, 2023. 2, 4, 5
- [57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022. 2, 3
- [58] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3
- [59] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023. 5
- [60] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023. 5
- [61] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3
- [62] Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985*, 2023. 3
- [63] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2