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

Imad Eddine Toubal[1,2*] Aditya Avinash[1] Neil Gordon Alldrin[1] Jan Dlabal[1] Wenlei Zhou[1]
Enming Luo[1] Otilia Stretcu[1] Hao Xiong[1] Chun-Ta Lu[1] Howard Zhou[1]
Ranjay Krishna[1,3 †] Ariel Fuxman[1] Tom Duerig[1]

[1]Google Research [2]University of Missouri [3]University of Washington
itdfh@umsystem.edu, {adity,nalldrin}@google.com

1. Concept names and descriptions

1.1. Agile Modeling dataset concepts

Arts and crafts: Image must contain arts and crafts.
Astronaut: Any picture that shows an astronaut, even if it's a drawing, clip art, etc. The astronaut should show clearly that they are associated with being an astronaut – usually indicated by a space suit or NASA jumpsuit.
Block tower: Image must contain a toy block tower made of legos or wood.
Dance: Photos of people dancing.
Emergency service: Image must contain emergency service, paramedics, firefighters, police, or rescue teams.
Gourmet tuna: Photos of gourmet dishes (i.e. fancy, elegant) that must contain tuna. This includes sushi, sashimi, seared tuna, a fancy ahi tuna salad. This does not include canned tuna, tuna sandwich, a photo of the fish tuna itself.
Hand pointing: A picture showing a hand pointing, with just the index finger extended. Does not include pictures of thumbs-up or pictures of hands with more than just the index finger extended. Picture with a straight finger pointing at or tapping a screen are included.
Hair coloring: Pictures that focus on people during the process of hair coloring or right after, before & after photos. Negatives: store front of hairdresser, boxes of dye.
Healthy dish: Photos of dishes with healthy food that is low in carbs
Home fragrance: Photos of any types of fragrances used for houses, including home perfumes, air fresheners for the house, scented candles, essential oils.
In ear headphones: Any headphones that are worn inside the ear, rather than covering it up. These types of headphones are inserted into the ear canal. As long as an in-ear headphone is in the picture, it is valid.
Pie chart: Any image with a pie chart, which is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.
Single sneaker on white background: Images depicting a single sneaker on a white or neutral-colored background (e.g beige). It can be a partial view of a sneaker (e.g. just the sole, or half of the sneaker is in view) but it cannot be just parts (e.g. just the shoe lace) . Negatives include images that have more than one shoe, that have different colored background, or a different style of shoe.
Stop sign: This includes photos of real-world, official stop signs. Imagine we would want to detect such stop signs for self-driving cars. Positives include any stop sign photos, including those temporary ones included in construction, or held in hand by a construction worker. If there's a stop sign on a banner or ads poster, even if it's in traffic, it would be a negative (we don't want the self-driving car to stop at that). Clip art or indoors stop sign are negative

1.2. Public dataset concepts

Hateful memes: Memes that are harmful, racist, or sexist

2. Search queries

The following is the set of search queries used to mine candidate images from the LAION [3] dataset during the data mining process of our system. All these search queries are generated using the LLM (PaLM-2 [1]) and encoded in joint CLIP [2] embedding space to retrieve candidate images.

Arts and crafts: craft room, crafts book, crafts for beginners, crafts tutorial, arts and crafts, craft store, crafts fair, craft project, crafts for sale, crafts, crafts for kids, art, craftsmanship, craft supplies, diy, crafts magazine, crafts for adults, handmade
Astronaut: astronaut, astronaut in space station module, astronaut in space gloves, astronaut in space boots, astronaut

*This work was done during an internship at Google.
†This work was done during working at Google.

in space flag, astronaut in space shuttle cockpit, astronaut in space suit, astronaut in orbit, astronaut in space backpack, astronaut in space station airlock, astronaut on moon, astronaut working in space, astronaut in space station, astronaut in space, astronaut in space helmet, astronaut in space shuttle, astronaut in space station cupola, astronaut walking in space

Block tower: tower of blocks, lego tower, tall lego tower, tall toy tower, tower of legos, tower of toys, towering wood, towering toys, towering blocks, block tower, towering legos, tall block tower, tower made of blocks, tall wooden tower, wooden tower, toy tower, tower of wood

Dance: street dance, flamenco, ballet, modern dance, bachata, ballroom dance, zouk, samba, people dancing, belly dance, salsa, merengue, dance performance, line dance, tap dance, hip hop, dancers, folk dance

Emergency service: police emergency, police officer at work, rescue boat, emergency response, police car, fire truck, rescue worker at work, emergency worker, medical emergency, firefighter, paramedic at work, rescue team, fire rescue, police, emergency service, rescue operation, firefighter at work, rescue helicopter, emergency vehicle, ambulance, paramedic

Gourmet tuna: tuna sushi, tuna sashimi, tuna salad, seared tuna, ahi tuna, gourmet tuna, tuna tartare, ahi tuna steak, tuna steak, fancy tuna

Hand pointing: hand pointing finger extended, hand pointing finger, hand pointing, hand pointing finger straight at them, hand pointing finger straight at someone, hand pointing finger straight at screen, hand index finger extended, hand pointing finger straight at something, hand pointing finger straight at person, hand pointing finger straight at me, hand pointing finger straight at us, hand pointing finger straight at you, hand pointing at screen, hand pointing finger straight at thing, hand pointing screen, hand pointing finger straight at object, hand pointing finger straight

Hair coloring: hair coloring, hair color salon, hair color before and after, hair color inspiration, hair color horror story, hair color mishap, hair color tutorial, hair color tips, hair dye, hair color stylist, hair color fail, hair color at home, hair color process, hair color mistake, hair color disaster, hair color gone wrong, hair color ideas, hair color, hair color gone bad

Healthy dish: healthy lunch, healthy sandwich, healthy dish, healthy burger, healthy meal, healthy food, low carb dish, healthy salad, healthy fish, healthy pizza, healthy dinner, healthy vegetarian, healthy vegan, healthy snack, healthy breakfast, healthy pasta, healthy chicken, healthy soup, healthy dessert

Home fragrance: home fragrance, home fragrance diffuser, scented candle, home scented candle, home scent, essential oil, home air freshener, air freshener, home essential oil, home scent diffuser, home room spray, home aroma diffuser, home smell diffuser, home perfume, home aroma, fragrance diffuser, home smell, room spray

In ear headphones: earphone, in ear headphones, in ear headphone, earbuds, in ear, headphone

Pie chart: pie chart maker, pie chart data, pie chart tutorial, pie chart percentage, pie chart illustration, pie chart design, pie chart template, pie chart chart, pie chart, pie chart infographic, pie chart graphic, pie chart diagram, pie chart creator, pie chart graph, pie chart example, pie chart generator, pie chart tool, pie chart software

Single sneaker on white background: sneaker on light beige background, sneaker on beige background, sneaker on neutral, sneaker on light background, sneaker on cream background, single sneaker, sneaker on white background, sneaker, sneaker on background, sneaker on solid background, sneaker on light off-white background, sneaker on light tan background, sneaker on neutral background, sneaker on light gray background, sneaker on light cream background, sneaker on plain background, sneaker on off-white background, shoe, sneaker on tan background, sneaker on white, sneaker on gray background

Stop sign: stop sign on road, stop sign on street, held stop sign, stop sign held by person, traffic stop sign, stop sign in traffic, stop sign in city, stop sign on interstate, stop sign, stop sign on highway, construction stop sign, stop sign in construction, real stop sign, stop sign on freeway, stop sign in rural area, official stop sign, stop sign in parking lot, stop sign in hand

3. LLM Prompts

We list a set of example prompts used in Modeling Collaborator Annotator below. When a description is unavailable for a given concept, we use the following prompt to auto-generate a structured description:

```
You are given a visual concept name.

Follow these steps:
<step1>You have to work as an expert
    ↪ linguist. There are some human
    ↪ annotators who need to determine
    ↪ if given images are in-scope or
    ↪ out-of-scope for this visual
    ↪ concept. Your task is to generate
    ↪  description of the visual
    ↪ concept which annotators can use
    ↪ to decide if any images are in-
    ↪ scope or out-of-scope for the
    ↪ given visual concept.</step1>
<step2>Provide an concept definition of
    ↪  this visual image in a few
    ↪ sentences.</step2>
```

```
<step3>Provide all the image attributes
    ↪  that an image must have in order
    ↪  to be in-scope for this visual
    ↪ concept.</step3>
<step4>Each attribute found in step2
    ↪ and step3 should be verbose,
    ↪ independent, self-explanatory and
    ↪  meaningful.</step4>
<step7>Write your response in following
    ↪  user friendly and readable
    ↪ format:
Visual concept definition:
<Add 2-3 line concept definition of the
    ↪  visual concept here.>

Image must have following attributes
    ↪ for it to be in-scope for this
    ↪ visual concept:
<Add details here as bullet points.>
</step7>
<visualConceptName>{CONCEPT_NAME}</
    ↪ visualConceptName>
```

The prompt for generating positive search queries based on a visual concept (used to fetch candidate positive images from an image database):

```
Your task is to help in finding
    ↪ positive (in-scope) images for a
    ↪ visual concept. You are given the
    ↪  name and the description of a
    ↪ visual concept. Description
    ↪ explains the attributes of an
    ↪ image that make it in-scope for
    ↪ this visual concept. It also
    ↪ explains the attributes of an
    ↪ image that make it out-of-scope
    ↪ for this visual concept.

Follow these steps:
<step1>List all the attributes of an
    ↪ image that make it in-scope for
    ↪ this visual concept.</step1>
<step2>Each attribute should be
    ↪ objective, complete, and self-
    ↪ explanatory.</step2>
<step3>Ensure that attributes you have
    ↪ found cover all the in-scope
    ↪ attributes or scenarios mentioned
    ↪  in the description. If not, add
    ↪ the missing in-scope attributes
    ↪ .</step3>
<step4>Based on all the in-scope
    ↪ attributes you have identified in
    ↪  step3, generate 20 Google Search
    ↪  keywords which can be used to do
    ↪  Google image search for finding
    ↪ diverse images with those in-
    ↪ scope attributes.</step4>
<step5>Ensure that your Google Search
    ↪ keywords cover all types of in-
    ↪ scope images mentioned in the
    ↪ description. If not, add Google
    ↪ Search keywords to find those
    ↪ types of in-scope images.</step5>
<step6>This is an important step. Some
    ↪ of the keywords you selected so
    ↪ far could be be suitable for
    ↪ searching out-of-scope images.
    ↪ Identify those keywords which are
    ↪  for searching out-of-scope image
    ↪ . Remove those Google Search
    ↪ keywords from your response.</
    ↪ step6>
<step7>Each search query should be 3-4
    ↪ words long, independent, self-
    ↪ explanatory and meaningful for
    ↪ internet image search.</step7>
<step8>If any of these queries are
    ↪ longer than 4 words, summarize
    ↪ them into 3-4 words.</step8>

<step9>Write your response in following
    ↪  xml format. Since your response
    ↪ will be programmatically parsed,
    ↪ your response should strictly
    ↪ follow this format:
'''xml
<google_search_keywords>
 <keyword></keyword>
 ...
</google_search_keywords>
'''
</step9>
<step10>Keep only xml in the response
    ↪ and remove other text.</step10>
<concept>{CONCEPT_NAME}</concept>
<description>{CONCEPT_DESCRIPTION}</
    ↪ description>
```

The prompt for generating negative search queries based on a visual concept (used to fetch hard negative images from an image database):

```
You have to work as an expert linguist.
    ↪   You are given a visual concept
    ↪ name and its description for the
    ↪ purpose of image classification.
```

```
Description might contains few carve-
    ↪ outs. Carve-outs are some special
    ↪  situations in which images
    ↪ should be classified as out-of-
    ↪ scope. Your task is to extract
    ↪ carve-out details from the
    ↪ description.

Follow these steps:
<step1>If the description does not
    ↪ contain any carve-outs, write
    ↪ your response in the following
    ↪ format and skip all of the
    ↪ following steps.
'''xml
<carveOutsInDescription>
  <carveOut>NOT_FOUND</carveOut>
</carveOutsInDescription>
'''
</step1>
<step2>If the description provides
    ↪ details on out-of-scope images
    ↪ for this visual concept, output
    ↪ the list of those carve-outs
    ↪ situations mentioned.</step2>
<step3>Output those in the following
    ↪ xml format. Since your response
    ↪ will be programmatically parsed,
    ↪ your response should strictly
    ↪ follow this format:
'''xml
<carveOutsInDescription>
  <carveOut></carveOut>
  ...
</carveOutsInDescription>
'''
</step3>
<step4>Keep only xml in the response
    ↪ and remove other text.</step4>

<concept>{CONCEPT_NAME}</concept>
<description>{CONCEPT_DESCRIPTION}</
    ↪ description>
```

The prompt template for generating a concept's positive attributes:

```
Your task is to understand the scope of
    ↪  a visual concept for image
    ↪ classification. You are given a
    ↪ visual concept name and its
    ↪ description.

Description explains the attributes of
    ↪ an image that make it in-scope
    ↪ for this visual concept. It also
    ↪ explains the attributes of an
    ↪ image that make it out-of-scope
    ↪ for this visual concept.

Follow these steps:

<step1>List all the attributes of an
    ↪ image that make it in-scope for
    ↪ this visual concept.</step1>

<step2>Each attribute should be
    ↪ objective, unambiguous, detailed,
    ↪  verbose and self-explanatory.</
    ↪ step2>

<step3>Check that attributes you have
    ↪ found cover all the positive
    ↪ attributes mentioned in the
    ↪ description. If not, add the
    ↪ missing attributes.</step3>

<step4>Write your response in following
    ↪  xml format. Since your  response
    ↪  will be programmatically parsed,
    ↪  your response should strictly
    ↪ follow this format:

'''xml
<positiveAttributes>
  <attribute></attribute>
  ...
</positiveAttributes>
'''
</step4>
<step5>Keep only xml in the response
    ↪ and remove other text.</step5>
<concept>{CONCEPT_NAME}</concept>
<description>{CONCEPT_DESCRIPTION}</
    ↪ description>
```

The prompt template for generating a concept's negative attributes:

```
You have to work as an expert linguist.
    ↪   You are given a visual concept
    ↪ name and its description for the
    ↪ purpose of image classification.
Description might contains few carve-
    ↪ outs. Carve-outs are some special
    ↪  situations in which images
    ↪ should be classified as out-of-
    ↪ scope. Your task is to extract
```

```
    ↪ carve-out details from the
    ↪ description.

Follow these steps:
<step1>If the description does not
    ↪ contain any carve-outs, write
    ↪ your response in the following
    ↪ format and skip all of the
    ↪ following steps.
'''xml
<carveOutsInDescription>
  <carveOut>NOT_FOUND</carveOut>
</carveOutsInDescription>
'''
</step1>
<step2>If the description provides
    ↪ details on out-of-scope images
    ↪ for this visual concept, output
    ↪ the list of those carve-outs
    ↪ situations mentioned.</step2>
<step3>Output those in the following
    ↪ xml format. Since your response
    ↪ will be programmatically parsed,
    ↪ your response should strictly
    ↪ follow this format:
'''xml
<carveOutsInDescription>
  <carveOut></carveOut>
  ...
</carveOutsInDescription>
'''
</step3>
<step4>Keep only xml in the response
    ↪ and remove other text.</step4>

<concept>{CONCEPT_NAME}</concept>
<description>{CONCEPT_DESCRIPTION}</
    ↪ description>
```

where CONCEPT_NAME and CONCEPT_DESCRIPTION are the subjective concept name and description.

For a final annotation decision for an image, we feed the following prompt to PaLM-2:

```
You are given the name and description
    ↪ of a visual concept. We showed
    ↪ the image to raters and asked
    ↪ many questions about the image
    ↪ and they gave the  answers.
    ↪ Questions and answers are also
    ↪ provided below. Your task is to
    ↪ answer some questions about the
    ↪ image. Follow these steps:

<step1>In the following steps, your
    ↪ text responses must be strictly
    ↪ based on the answers provided in
    ↪ raters' responses.</step1>

<step2>Provide the out-of-scope
    ↪ attributes present in the image
    ↪ .</step2>

<step3>Provide the in-scope attributes
    ↪ present in the image.</step3>

<step4>Provide the the in-scope
    ↪ attributes missing in the image
    ↪ .</step4>

<step5>Classify the image based on the
    ↪ following rules. Rules must be
    ↪ followed in the given order.

<classificationRules>

<rule1>If  the image has EVEN ONE of
    ↪ the out-of-scope attributes, it
    ↪ must be classified negative for
    ↪ this visual concept.</rule1>

<rule2>The image must have all the
    ↪ required positive attributes to
    ↪ classify the image as positive
    ↪ for this visual concept.   If
    ↪ image has all the required
    ↪ positive attributes, classify the
    ↪  image as positive. Otherwise
    ↪ classify it as negative.</rule2>

<rule3>In all other cases, classify the
    ↪  image as negative.</rule3>
</classificationRules></step5>

<step6>Add following details to your
    ↪ response strictly in this format:
Decision: "Positive" or "Negative"
Reasons: <Provide list of reasons why
    ↪ this image is Positive or
    ↪ Negative> </step6>

<step7>Make sure your response only
    ↪ contains text and no python code
    ↪ .</step7>

<concept>{CONCEPT_NAME}</concept>
<conceptDescription>{
```

```
    ↪ CONCEPT_DESCRIPTION}</
    ↪ conceptDescription>
<raterResponses>{
    ↪ PALI_QUESTIONS_AND_ANSWERS}</
    ↪ raterResponses>
```

where `PALI_QUESTIONS_AND_ANSWERS` is a formatted string of questions fed to PaLI-X VQA and their respective answers.

4. Ablations

To measure the effect of the expert involvement we show Fig. 2. Overall, expert collaboration improves the performance of the distilled model. As the number of expert-labeled examples increases (0 to 2000 out of total training 4000 examples), the recall, F1, and auPR scores of the model also increase.

To show the impact of additional automatically annotated data on the performance of the final output model on easy vs hard concepts, we show Fig. 1.

5. Annotator Configurations

We define the following settings for the Annotator:

A. `use_positive_attributes_for_questions`: Whether to generate positive questions from the attributes or directly from the concept description.

B. `generate_negative_questions`: Whether or not to generate negative questions. Sometimes these questions result in over-predicting negative classes.

C. `use_captioning_questions`: Whether to use a captioning VLM to generate a detailed description of the image

D. `generate_fixed_num_of_questions`: Fix the number of questions instead of having the LLM generate as many questions as possible.

E. `final_rating_without_attributes`: Whether to use negative and positive attribute in the final annotation stage.

Using a grid search, we use different configurations for different concept as described in Tab. 1.

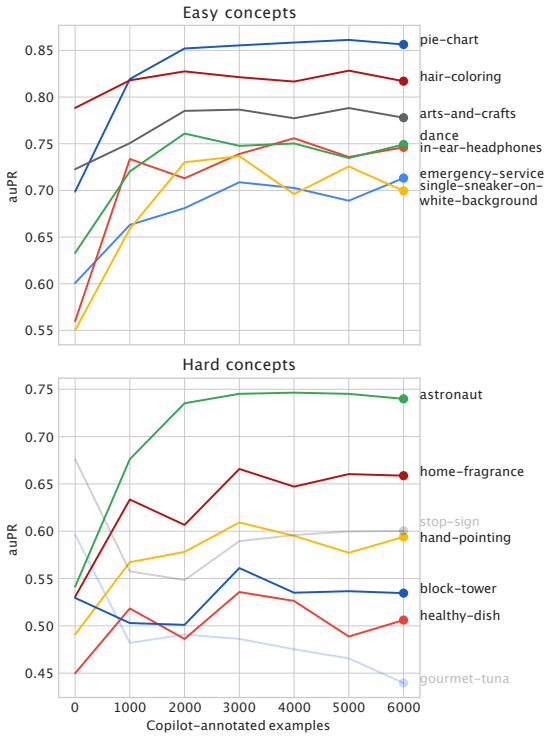

Figure 1. The impact of adding additional automatically annotated images on the final model quality (using the auPR metric). 100 user-annotated examples are used in addition to the thousands of Modeling Collaborator examples.

	Configuration				
Concept	**A**	**B**	**C**	**D**	**E**
arts-and-crafts				✓	✓
astronaut			✓	✓	✓
block-tower			✓	✓	✓
dance			✓	✓	✓
emergency-service					✓
gourmet-tuna				✓	✓
hair-coloring					✓
hand-pointing					✓
healthy-dish			✓	✓	✓
home-fragrance				✓	✓
in-ear-headphones			✓	✓	✓
pie-chart			✓	✓	✓
single-sneaker				✓	✓
stop-sign				✓	✓

Table 1. Configuration settings used for each concept of the Agile Modeling [4] dataset.

References

[1] 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. 1

[2] 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. 1

[3] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes,

Figure 2. The impact of Modeling Collaborator and expert collaboration on the performance of the distilled model. 4,000 total training examples were used per concept. The x-axis represents how many of those examples were labeled by the expert (concept owner), ranging from no examples (0%) to 2,000 examples (50%).

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. 1

[4] 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. 6