A. Limitations

Here we discuss a few limitations of our study and how these can be addressed in future work. Most of these are caused by the extensive costs required to conduct more user studies, as explained below.

User study scale. Our evaluation involved nine participants across 18 sessions (approximately 27 hours of study time), which may limit the statistical power and generalizability of our findings. Each session required extended interaction with the system to iteratively refine subjective concept definitions, leading us to prioritize depth of interaction over study scale. The total study time is comparable to prior work at the intersection of HCI and ML systems evaluating subjective, contextual workflows, which often report 10–30 hours of user study time [17, 45]. Future work should examine larger participant pools and a broader set of concepts to strengthen external validity.

Component-level ablations. Our evaluation does not isolate the contributions of individual components within Agile Deliberation, such as borderline image retrieval and prompt optimization. Instead, we evaluate the system as an integrated workflow. This design reflects how these mechanisms operate together in practice: formative interviews with domain experts suggested that identifying informative borderline examples and refining definitions are tightly coupled activities in concept deliberation. Future work should conduct additional user studies to independently examine the effects of these components.

Evaluation across generative models. Our implementation of Agile Deliberation primarily relies on a single generative model (Gemini 2.5 Flash), and we did not evaluate the full pipeline across other generative models. On one hand, Agile Deliberation only requires inference access and can in principle be used with any API-accessible VLM, including newer models with stronger reasoning capabilities. Future work should evaluate the framework with a broader range of models to better understand how model choice affects system performance and usability.

Beyond the generalizability of the full pipeline, there is also the question of whether concept definitions produced by Agile Deliberation transfer effectively to other VLMs. We provide a preliminary investigation of this question in the next section. Future work should more systematically evaluate the transferability of concept definitions across a broader range of models.

B. Transferability of Concept Definitions Across VLMs

We conducted an additional experiment to examine whether concept definitions produced by Agile Deliberation transfer to other VLMs. Specifically, we took concept definitions produced under both the Agile Deliberation condition and the baseline condition—both implemented using *Gemini 2.5 Flash*—and applied them to external, smaller VLMs for classification: *Qwen3-VL-8B* and *Qwen3-VL-30B-A3B-Instruct* [40]. We presented the results at the Table 3.

Overall, the results suggest that concept definitions from both conditions transfer to other models, but Agile Deliberation produces definitions that are more sensitive to model capability. Definitions from the Manual condition remain relatively similar across models, while Agile Deliberation definitions continue to show benefits on the Qwen models, with the strongest gains appearing on more capable models. This pattern suggests that Agile Deliberation yields richer and more complex concept definitions, whose advantages become more visible as model reasoning ability improves.

It is also worth emphasizing that these improvements are obtained despite the fact that Agile Deliberation optimized the concept definition for a different model (i.e. *Gemini 2.5 Flash*), and we expect even bigger gains when the target model is also the one used as classifier in the Agile Deliberation process.

C. Interview Analysis

C.1. Concept Deliberation by Experts

We first conducted a qualitative analysis of 20 concept definitions created by professional content moderators as part of their regular workflow. These materials included both the finalized definitions and their accompanying discussion threads, giving us visibility into how teams deliberate over subjective visual concepts. We learned that practitioners typically begin by *scoping* their concept definitions and then iteratively *refining* them by searching for and reflecting on borderline images. Both of these stages require domain expertise and newcomers in the team need to spend significant time familiarizing themselves with this process.

Stage 1: Concept Scoping — Structured definitions facilitate concept deliberation. Experts often structured their initial definitions by decomposing composite concepts into simpler unit components. For example, a composite idea such as *before-and-after transformations for achievements* was separated into the visual pattern of a *before–after layout* and the semantic notion of an *achievement*, whereas a concept like *beautiful images* was treated as a single unit. Each unit concept was then expanded into positive and negative visual categories—for instance, *city parks* versus *industrial factories* when reasoning about *beautiful images*. This hierarchical and contrastive structure helped practitioners articulate core visual signals and establish the initial decision boundary.

Stage 2: Concept Iteration — Borderline images are essential for iterative refinement. Borderline images then played a critical role in refining these scoped definitions.

	Model:**Qwen3-VL-8B**	Participants in **Agile Deliberation Condition**			Participants in **Manual Condition Condition**		
	Condition	F1	Precision	Recall	F1	Precision	Recall
Paid to play	Zero-shot	0.46 (.09)	0.31 (.08)	**0.95** (.07)	0.36 (.06)	0.22 (.05)	**0.93** (.06)
	Assigned Deliberation System	**0.52** (.07)	**0.44** (.10)	0.68 (.15)	**0.51** (.06)	**0.39** (.08)	0.76 (.13)
	Δ	6%	13%	-27%	15%	17%	-17%
Healthy food	Zero-shot	0.47 (.14)	0.32 (.13)	**0.96** (.03)	**0.81** (.04)	0.73 (.06)	**0.91** (.02)
	Assigned Deliberation System	**0.52** (.15)	**0.38** (.14)	0.91 (.08)	**0.81** (.03)	**0.80** (.04)	0.83 (.03)
	Δ	5%	6%	-5%	0%	7%	-8%

	Model:**Qwen3-VL-30B-A3B-Instruct**	Participants in **Agile Deliberation Condition**			Participants in **Manual Condition Condition**		
	Condition	F1	Precision	Recall	F1	Precision	Recall
Paid to play	Zero-shot	0.49 (.09)	0.34 (.08)	**0.97** (.05)	0.38 (.09)	0.24 (.06)	**0.93** (.08)
	Assigned Deliberation System	**0.59** (.09)	**0.50** (.12)	0.75 (.12)	**0.50** (.09)	**0.39** (.13)	0.78 (.13)
	Δ	10%	16%	-22%	12%	15%	-15%
Healthy food	Zero-shot	0.47 (.14)	0.32 (.13)	**0.95** (.03)	**0.83** (.03)	0.75 (.05)	**0.88** (.10)
	Assigned Deliberation System	**0.52** (.19)	**0.40** (.18)	0.83 (.21)	0.78 (.04)	**0.79** (.04)	0.78 (.05)
	Δ	5%	8%	-12%	-5%	4%	-10%

Table 3. **Transfer performance of concept definitions across participant conditions on additional VLMs.** F_1, precision, and recall are averaged over participants in each condition (standard deviation in parentheses). We apply concept definitions produced under the Agile Deliberation and Manual conditions using *Gemini 2.5 Flash* to additional VLMs to examine how well these definitions transfer across models. "Assigned Deliberation System" denotes Agile or Manual per condition. Since participants varied in the complexity of their understanding for the same concept, classification performances are not directly comparable across these two groups. Instead, we focus on each system's improvement over its respective zero-shot baseline (indicated by Δ).

Such images revealed ambiguities that the initial structure did not address—for example, encountering an industrial site that had been remodeled into a park prompted practitioners to reconsider how transitional or mixed scenes should be classified under *beautiful images*. These borderline cases made gaps in the definition immediately visible and helped practitioners clarify ambiguous language, resolve inconsistencies, and sharpen the intended scope. In practice, visual inspection of borderline images proved far more intuitive than abstract reasoning alone, providing concrete signals that guided iterative refinement.

C.2. Challenges in Concept Deliberation

Our analysis of 20 expert-authored concept definitions gave us an initial picture of how subjective concepts are scoped and refined. Building on this insight, we conducted five formative studies with professional content moderators to better understand the challenges they encounter. Each session was one hour long and consisted of two parts. In the first part, participants defined a new visual concept using a provided template. They scoped the concept definition, searched for borderline images in a think-aloud manner. In the second half, we conducted a semi-structured interview to discuss the challenges they encountered, how they searched for borderline images, and how they expressed nuanced decision boundaries when preparing a concept for an LLM-based classifier.

Even with substantial expertise, participants still reported several difficulties. They spent a large amount of time searching for edge-case images that could clarify the subtle boundaries of subjective concepts. Although initial scoping was usually straightforward, identifying truly borderline images was much harder. It required them to question their own intuitions, and locating such cases in a large dataset was slow and often frustrating. Existing tools provided only limited support. Similarity search tended to surface images that were close in embedding space but not

genuinely borderline, so these results offered little insight into conceptual edges. Participants also tried using LLMs to generate search queries, but they found that LLMs often lacked awareness of the visual complexity in real datasets. Many borderline cases involved cropped, blurred, zoomed-in, or partially occluded images, or images embedded in cluttered backgrounds. These subtleties were difficult for LLMs to reason about [21, 46]. As a result, experts still relied heavily on manual exploration to identify the borderline cases that were most helpful for refining their definitions.

D. Interface of Agile Deliberation

Figure 4–7 show screenshots from the interactive notebook interface used in our study. The interface supports the two stages of Agile Deliberation: *concept scoping*, where users review candidate subconcepts derived from an initial concept description, and *concept iteration*, where users inspect and label borderline images to refine the concept definition over multiple rounds.

E. LLM Prompts

E.1. Image Classifiers

Prompts that classify images based on the structured definition. We convert the resulting ratings into binary labels by thresholding at 3.

```
<role>You are an expert image annotator.</role>
<input>You will be provided with a single image, an
image caption and a concept definition.</input>

<task>
- Thoroughly examine the image.
- Carefully consider all details of the concept
- Determine if the image satisfies every aspect of the
 given concept.
</task>

<step1>
Explicitly write down the specific requirements this
concept demands. Make sure your decomposition is
equivalent to the original definition. There are two
special cases that demand your attention:
- If the concept is defined by a list of necessary
conditions, you should start by checking each
condition explicitly. You should only give a high
rating if all conditions are satisfied.
- If the concept is defined by a list of positive and
negative conditions, you should start by checking each
 condition explicitly. You should give a high rating
if any of the positive conditions are satisfied and
none of the negative conditions are satisfied.
- A concept can be first defined by a list of
necessary conditions, and then each necessary
condition can be further defined by a list of positive
 and negative conditions. You should follow the same
logic recursively to check if the image satisfies the
condition.
</step1>

<step2>
Based on responses in the first step, for each
condition, you should determine whether the image
satisfies the condition.
</step2>

<step3>
You should then combine your responses for individual
conditions to determine whether the image satisfies
the overall concept or not.
Your reasoning should be based solely on the
definition, the image, and the image caption. Do not
include any information that is not provided (no
hallucinations).
</step3>

<step4>
Based on your reasoning in the first step, determine
whether the image completely fulfills every aspect of
the concept.
You should rate how in-scope the image is on a 1-5
Likert Scale where
- Rate 5 if the image fully aligns with the concept
- Rate 4 if the image mostly aligns with the concept;
the small problem is a result of small ambiguities in
the definition or small visual complexities in the
image
- Rate 3 if there are no strong evidence that indicate
 that the image violates the concept, but the
supporting evidence is also not strong enough to
support a rating of 4 or 5.
- Rate 2 if the image violates some parts of the
concept but there are some elements that are relevant
to the concept.
- Rate 1 if the image does not align with the concept
description at all.
</step4>

<step5>
Based on your answer in previous steps, provide a one-
sentence summary why you give this answer.
</step5>

Provide your answer in the following XML structure:
<requirements>Explicitly write down the specific
requirements this concept demands.</requirements>
<condition-eval>Write down your evaluation for each
condition at step2 here.</condition-eval>
<evaluation>You should differentiate between concept
definitions that requires the satisfaction of all
conditions and those that only require the
satisfaction of one of the conditions. Describe your
evaluation reasonings in the step4 here</evaluation>
<decision>Rate on a 1-5 Likert Scale where 5 means the
 image is fully in-scope and 1 means the image is
fully out-of-scope.</decision>
<summary>Provide your summary in the step5 here</
summary>
The output will be later wrapped in a <root> tag, so
do not wrap the content above in any tag such as xml,
or root in your output.
```

E.2. Decomposition module

Prompts that decompose a composite concept

```
<role>You are an expert linguist recognized
internationally.</role>
<input>You will receive a visual concept name and a
description.</input>
<task>There are human image annotators who need to
determine if an image is in scope or out of scope of
this visual concept. Your job is to break down this
visual concept into at most two necessary conditions:
the conjunction of these necessary conditions will be
logically equivalent to the given visual concept. In
other words, images that satisfy all these conditions
will be exactly images that satisfy the visual concept
. We call this process as decomposition. The final
```

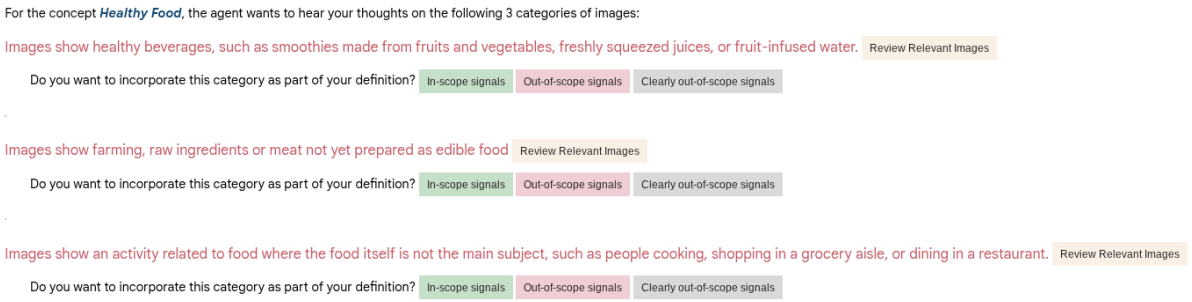

Figure 4. **Concept Scoping interface.** The system proposes candidate positive and negative subconcepts derived from the user's initial concept description. For each subconcept, representative images retrieved from the dataset are shown for inspection. The user decides whether the subconcept should be included in the structured concept definition.

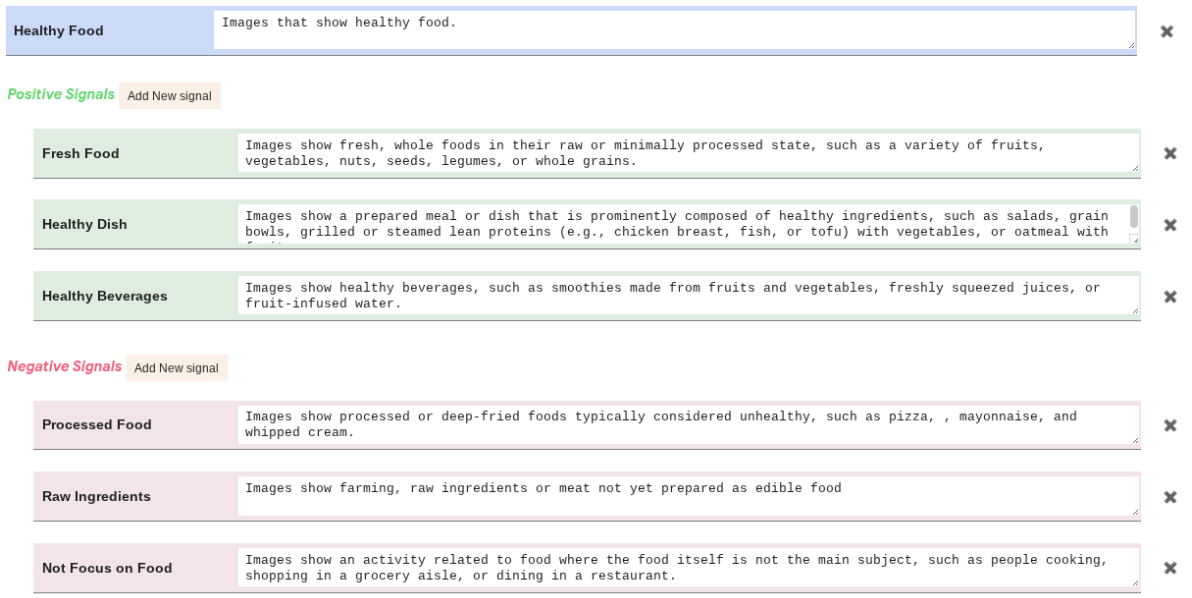

Figure 5. **Structured concept definition after Concept Scoping.** Based on the user's selections, the system constructs an initial concept definition composed of positive and negative subconcepts that capture the core visual signals of the concept.

```
goal is to make it easier and more accurate for human
raters to determine if an image is in-scope or out-of-
scope for the given visual concept.
</task>


<step1>
Rewrite the description by removing words that are
redundant, too specific, or actually not necessary for
 this overall concept. Other than removing words, you
should keep most parts of the description intact. This
 is because users might provide a draft description in
 the beginning without knowing what are in-scope
images look like in the wild. As a result, they might
introduce too many details that are not actually not
necessary for the concept.
</step1>

<step2>
Based on your response in step1, reason how would you
decompose the visual concept into several conditions.
Each necessary condition should be concise, self
```

```
explanatory, and easily understood by human annotators
. Each necessary condition should only focus on one
concept, and you should not generate a still complex
necessary condition. For each condition, you should
provide a description of the condition, and a concept
name that summarizes the aspect this condition focuses
 on. Explicitly write down a description and a name
for your decomposed conditions here. Remember not all
concepts can be further decomposed.
</step2>

<step3>
Examine the following aspect for your decomposition.
(1) Each condition must not significantly overlap with
 other conditions in their focused concepts.
Examples
- the condition that "a family is gathering together"
and "a family is having a meal together" are too
similar to each other despite they have different
focus.
- the condition that "a vase is broken into pieces"
and "someone used some tools to break a vase into
```

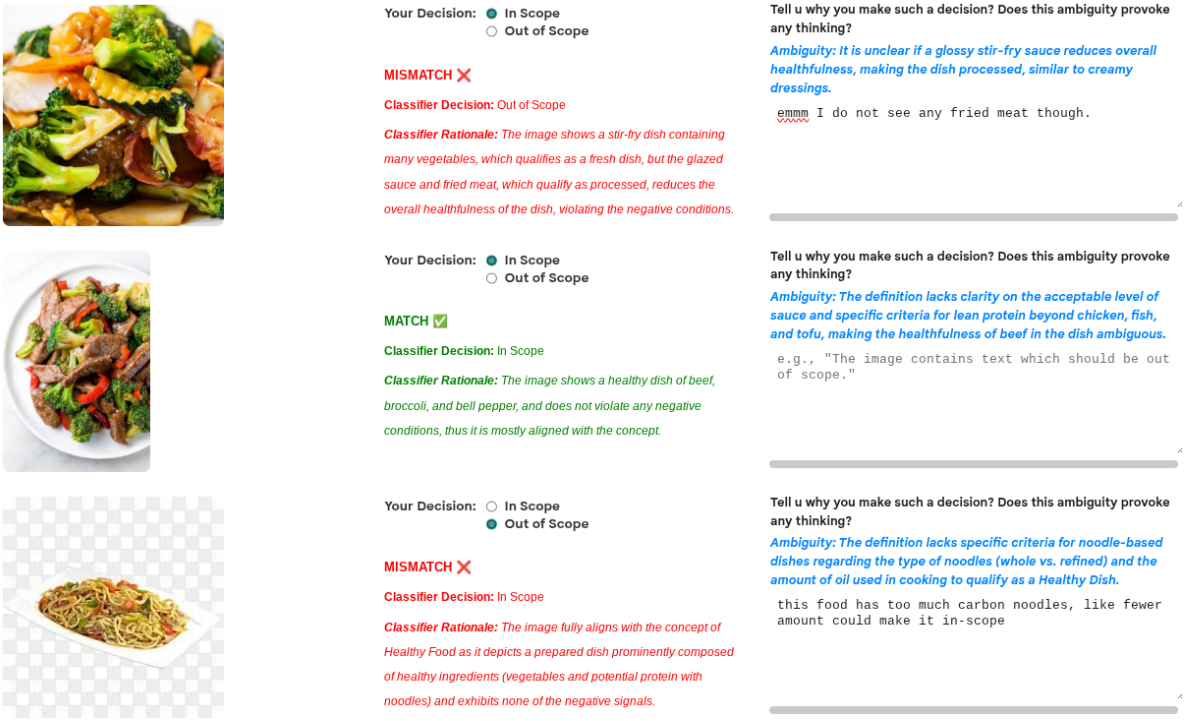

Figure 6. **Concept Iteration interface.** The system retrieves a batch of borderline images that are semantically ambiguous under the current concept definition. The user reviews the model's prediction and explanation and then labels each image as in-scope or out-of-scope.

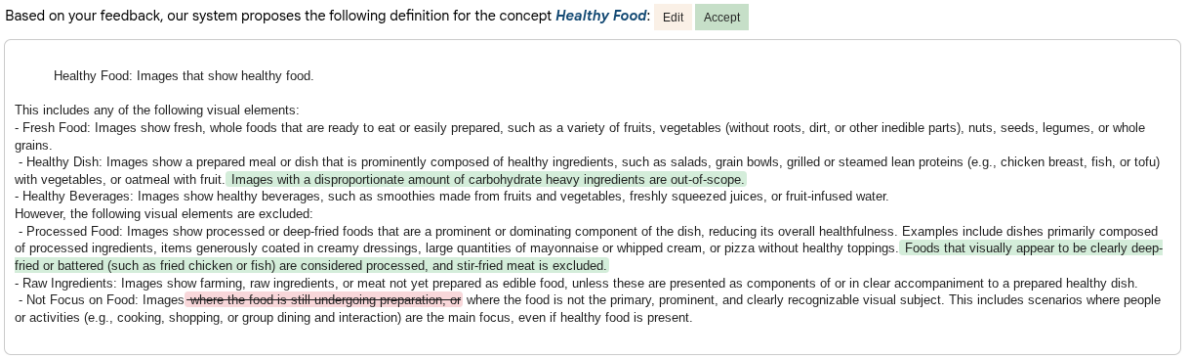

Figure 7. **Updated concept definition after one iteration round.** User feedback on borderline images is incorporated to refine the concept definition, producing an improved classifier aligned with the user's interpretation.

```
pieces" are too similar to each other despite slightly
 different wording and focus.."
(2) Each condition carries meaningful information,
meaning it should not hold true for every image.
For instance, the condition ``the image describes an
object'' is too broad and would be true for all images
.
</step3>
<step4>
Write down your decomposed conditions.
In cases when you find it hard to decompose the
concept, you can just write down a improved concept
description. of the original concept here.
```

```
You should follow these guidance in your description.
(1) Avoid using verbs, adjectives, or adverbs that
carry nuances unless they are an important part of the
 concept.
Examples
- if the original concept only uses the phrase "show a
 group of children", then avoid adding using the
phrase "depict a group of children" as "depicts"
introduces the slight emphasis on visual aspects and
ignore the possibility of textual information in the
image.
- if the original concept only mentions the phrase "
electronic devices", then avoiding using the phrase "
such as a phone or a laptop" as the new listed
```

```
examples suggest a focus on these specific examples.
- if the original concept only uses the phrase "show a
 beautiful park", then avoid using the phrase "clearly
 show a beautiful park" because "clearly" implies a
degree of visibility to the original concept.

(2) Avoid further defining complex, abstract, or
subjective concepts in the original concept
definitions.
We only want to break down a composite concept into
more unit concepts, and we will define them more
clearly in the next round.
Examples:
- if the original concept is "show a beautiful
painting", you should just decompose it into "show a
painting" and "the painting is beautiful"; you do not
need to explicate what make a painting beautiful.
- if the original concept is "show people are gathered
 happily", you should just decompose it into "show
people" and "people are gathered happily"; you do not
need to explain visual elements of being happy.

(3) You must not include information beyond the
provided information, as new information would
effectively change the intended meaning of the concept
.
Examples
- if the original concept "woman in a bra" does not
mention the context "at the beach", then do not add
this context in your necessary conditions.
</step4>

Provide your answer in the following XML structure:
<new-description>Your refined description</new-
description>
<reasoning>Add your reasoning here at step2</reasoning
>
<examination>Add your examination here at step3</
examination>
<conditions>
  <condition>
    <description>Add a condition here</description>
    <name>a short name that summarizes the description
    </name>
  </condition>
  <condition>
    <description></description>
    <name></name>
  </condition>
  <!-- Add more necessary conditions here if needed.
  -->
</conditions>

[omit few-shot examples here]

  <visualConceptName>{definition.concept}</
  visualConceptName>
  <visualConceptDescription>{definition.description}</
  visualConceptDescription>
```

Prompts that brainstorm candidate categories of a given concept

```
<role>You are an expert linguist recognized
internationally.</role>
<input>You will be provided an overall concept
definition and a focus concept.</input>
<task>
The concept owner wants to catch all images that are
in-scope for the focus concept.
To make sure that his decisions about image
classifications are consistent, he wants to explicitly
 clarify the decision-making boundaries for this focus
 concept.
However, as he starts with a few images, it is
possible that he either starts with a narrower concept
 or a broader concept than he actually wants.
Your task is to infer what are the golden subconcepts
that the concept owner wants to include within the
scope of this focus concept.
</task>

<step1>
Reason what is the primary concept that the concept
owner wants to explicitly define.
There are a few requirements.
1) This description might mention several concepts but
 you should only focus on the primary concept.
2) If the context indicates that the focus concept is
part of the necessary signals of a larger concept,
then the primary concepts of these necessary signals
should focus on different subconcepts of this larger
concept.
In other words, your primary concept should have a
different focus than those of the other necessary
signals.
3) The primary concept should be more categorical (
concepts where you could think about specific
instances) rather than descriptive (where you could
only describe different aspects of the concept).
Examples of categorical concepts are "fruit", '
electronic devices', 'physical affection', 'outdoor
activities',
whereas examples of descriptive concepts are "sleeping
 person", 'romantic relationship', 'sexual suggestive
content'.
</step1>

<step2>
List out categories of subconcepts that have been
explored before.
Here the categories mean a way to categorize specific
subconcepts, for instance, 'earphones' as a subconcept
 can belong to the category 'electronic devices' and '
accessories' at the same time.
This includes the following two cases:
1) categories that are already included in the concept
 definition as positive or negative signals.
2) categories that have beed explored in the previous
rounds of brainstorming.
</step2>

<step3>
Based on your answer in step2 and step3, reason and
propose a category of subconcepts that you think is
the most coherent and widely recognized.
While you can include previous explored subconcepts,
your category should not significantly overlap with
previously explored categories that you listed in
step3
This is because a subconcept can belong to multiple
categories at the same time.
In particular, we have the following requirements:
<requirements>
1. You should ensure that this category itself is a
well-defined and well-known concept so that average
people can easily tell whether an image satisfies this
 category or not.
2. Your category should not be too narrow that it only
 covers one or two instances.
3. In cases where there are many potential categories
of subconcepts, you should prioritize the one that
most people would agree to be in-scope for the concept
.
4. You do not aim for proposing a category that
includes the most subconcepts; Instead, you should
prioritize proposing a category that is coherent, and
well-defined.
</requirements>

<examples>
- For the primary concept "fruits", "fruits with red
internal flesh" is not a well-known concept, whereas "
```

```
citrus fruits" is.
- For the primary concept "flowers", "Flowers with
five petals" is not a well-known concept and also
difficult to recognize in an image; "Rose varieties"
is instantly recognizable.
- For the primary concept "birds", "birds of prey" is
a well-known concept, visually distinct, and specific
without being too narrow.
- For the primary concept "buildings," while "
buildings with green roofs" may not be a widely
recognized concept, "religious buildings" is a well-
known concept that is often visually distinctive.
- For the primary concept "health care products,"
while "Vitamin C supplements" might seem like a well-
defined category, it is too narrow.
  "Dietary supplements" is a more suitable category
  encompassing various products like vitamins,
  minerals, herbs, and fish oil, providing a broader
  yet well-defined and easily recognizable concept.
</examples>
</step3>

<step4>
For the category in step4, you should write a one-
sentence description and a shortname.
Your description should be concise, self explanatory,
and easy for an average person to determine if an
image satisfies this subconcept or not.

<description-requirements>
You should be careful about your language, in
particular, there are several requirements.
1) The recommended format for the description would be
 "Images show [a general term for the subconcept],
such as [at most three specific examples from step3]".
These examples shoulld be representative of the
subconcept and should be as specific as possible so
that human image annotators can easily know whether an
 image includes this example or not.
These examples should form a coherent category, and
should be specific.
Your short name for the description should be exactly
the term that covers this set of examples.

2) Avoid concept descriptions with too many specific
and unnecessary details.
   e.g., for the concept 'beverages', your subconcept
   description should just be 'Images showing various
   types of tea drinks such as green tea, black tea,
   and herbal tea'
   rather than 'Images that show people drinking
   various types of tea drinks with different colors
   and flavors such as green tea, black tea, and
   herbal tea'
  Similarly, avoid adding too many unnecessary details
   to the examples you provide.
  For instance, "eagles" is a good example of the
  category "birds of prey", but "bald eagles" or "
  eagles in the sky" are not.

3) Be careful about your word choices of verbs, nouns,
 or adjectives, which might carry unexpected nuances.
   e.g., be careful about using 'depict' or 'mention',
    or 'show' as the previous two verbs introduce the
   slight emphasis on visual or textual aspects.
   e.g., be careful about using adjectives like '
   clearly' or 'explicitly' as they might suggest a
   degree of visibility to the original concept.
</description-requirements>
</step4>

Write your output strictly in this valid xml format:
  <repeat-focus-concept>Repeat the focus concept here
  .</repeat-focus-concept>
  <primary-concept>List out the primary concept this
  categorical concept focuses on here.</primary-
  concept>
  <explored-subconcepts>
  List out the questions that have been asked for this
   concept before if any in a bullet point list at
  step2.
  </explored-subconcepts>
  <category-reasoning>
  List out your reasonings at step3 here.
  </category-reasoning>
  <subconcept>
    <description></description>
    <name></name>
  </subconcept>

[omit few-shot examples here]

<conceptDefinition>
  {str(definition)}
</conceptDefinition>
<previous-signals>
  {previous_signal_str}
</previous-signals>s
<context>{context}</context>
```

Prompts that brainstorm more borderline categories of a given concept

```
<role>You are an expert linguist recognized
internationally.</role>
<input>You will be provided a concept definition and
an optional context for this concept.</input>
<task>
The concept owner wants to catch all images that are
in-scope for this concept.
To make sure that his decisions about image
classifications are consistent, he wants to explicitly
 clarify the decision-making boundaries for this
concept.
However, as he starts with a few images, it is
possible that he either starts with a narrower concept
.
Your task is to suggest a new borderline subconcept
that the concept owner might want to include.
</task>

<step1>
Reason what is the primary concept that the concept
owner wants to explicitly define.
There are a few requirements.
1) This description might mention several concepts but
 you should only focus on the primary concept.
2) If the context indicates that the focus concept is
part of the necessary signals of a larger concept,
then the primary concepts of these necessary signals
should focus on different subconcepts of this larger
concept.
In other words, your primary concept should have a
different focus than those of the other necessary
signals.
3) The primary concept should be more categorical (
concepts where you could think about specific
instances) rather than descriptive (where you could
only describe different aspects of the concept).
Examples of categorical concepts are "fruit", '
electronic devices', 'physical affection', 'outdoor
activities',
whereas examples of descriptive concepts are "sleeping
 person", 'romantic relationship', 'sexual suggestive
content'.
</step1>

<step2>
If the user starts with a narrower concept, reason
what will be the broader concept that the user might
want to define.
You should examine the given context and summarize the
 broader concept the user might intend to say in
```

```
replacement of the primary concept.
This broader concept should fit nicely with other
parts of the context.

<example>For the concept 'health supplements' within
the context of "images that show health supplements to
 promote wellness", the broader concept might be '
wellness products'</example>
<example>For the concept 'fake websites' within the
context of "images that show fake websites as part of
online fraud", the broader concept might be 'online
fraudulent schemes'</example>
<example>For the concept 'electronic devices' within
the context of "images that show electronic devices in
 a library", the broader concept might be 'things we
can find in a library'</example>
</step2>

<step3>
List out categories of edgecase categories that have
been explored before from the previous signals input.
</step3>

<step4>
Based on your answer in step1 and step2, what other
subconcepts might be in-scope for this broader concept
 in step2 but obviously not part of the primary
concept in step1.
In other words, you should not focus on detailing
specific edgecase categories of this primary concept.
Instead you should think about what other subconcepts
(replacing the primary concept) frequently appear in
the same context.

There are a few requirements.
<requirements>
1) You should NOT focus on detailing specific edgecase
 categories of this primary concept.
2) Your category should NOT significantly overlap with
 the subconcepts that have been explored at step3.
3) Your category should not refer to examples that
significantly overlap with the examples that have been
 explored before in step2.
4) We will later define the other necessary signals
for this concept, so your category should NOT try to
define other necessary signals.
</requirements>

<example>For the concept 'health supplements' within
the context of "images that show health supplements to
 promote wellness", 'fresh fruits', 'yoga mats', or '
spa treatments' might also be interesting because they
 can also visually symbolize natural and holistic
approaches to health and well-being.</example>
<example>For the concept 'fake websites' within the
context of "images that show fake websites as part of
online fraud", "counterfeit product pages", "fake
social media accounts", or "fraudulent payment forms"
might also be interesting because they can also be
depicted as part of online fraud schemes.</example>
<example>For the concept 'electronic devices' within
the context of "images that show electronic devices in
 a library", 'board games', 'books', or 'musical
instruments' might also be interesting because they
might also appear in a library despite not electronic
.</example>
</step4>

<step4>
Based on your reasoning in step4, write a one-sentence
 description and a shortname for your borderline
subconcepts.

<description-requirements>
You should be careful about your language, in
particular, there are several requirements.
1) The recommended format for the description would be
 "Images show [a general term for the subconcept],
such as [at most three specific examples from step3]".
These examples shoulld be representative of the
subconcept and should be as specific as possible so
that human image annotators can easily know whether an
 image includes this example or not.
These examples should form a coherent category, and
should be specific.
Your short name for the description should be exactly
the term that covers this set of examples.

2) Avoid concept descriptions with too many specific
and unnecessary details.
   e.g., for the concept 'beverages', your subconcept
   description should just be 'Images showing various
   types of tea drinks such as green tea, black tea,
   and herbal tea'
   rather than 'Images that show people drinking
   various types of tea drinks with different colors
   and flavors such as green tea, black tea, and
   herbal tea'
  Similarly, avoid adding too many unnecessary details
   to the examples you provide.
  For instance, "eagles" is a good example of the
  category "birds of prey", but "bald eagles" or "
  eagles in the sky" are not.

3) Be careful about your word choices of verbs, nouns,
 or adjectives, which might carry unexpected nuances.
   e.g., be careful about using 'depict' or 'mention',
    or 'show' as the previous two verbs introduce the
   slight emphasis on visual or textual aspects.
   e.g., be careful about using adjectives like '
   clearly' or 'explicitly' as they might suggest a
   degree of visibility to the original concept.
</description-requirements>
</step4>

Write your output strictly in this valid xml format:
  <primary-concept>List out the primary concept this
  categorical concept focuses on here.</primary-
  concept>
  <broader-concept>List out the broader concept that
  the user might want to define here.</broader-concept
  >
  <previous-signals>List out the previous signals
  explored before here at step3 in a bullet point list
  .</previous-signals>
  <reasoning>
  List out your reasonings at step4 here.
  </reasoning>
  <subconcept>
    <description></description>
    <name></name>
  </subconcept>

<conceptDefinition>
  {str(definition)}
</conceptDefinition>
<previous-signals>
{previous_signal_str}
</previous-signals>
<context>{context}</context>
```

E.3. Borderline Image Retrieval Module

Prompts that generate borderline/in-scope descriptions of the subjective concept

```
<role>You are an expert linguist who is good at
brainstorming creatively.</role>
<input>
You are given a structured concept definition and a
list of previously generated descriptions.
</input>
<task>
```

```
Your task is to generate {num_descriptions} more
description that covers a possibly {image_type}
category of images.
This category of images should be different from the
categories covered by previous descriptions.
Since users still explore and improve their concept
definition, you should also consider similar
categories
that might be fully covered by the current definition
but are relevant.
This generated description will later be used to find
images that satisfy the concept through a search
engine.
</task>

<define-a-concept>
Before answering the question, it is a must for you to
 understand how we define a concept in a structured
and iterative way.
- If the concept is defined by a list of necessary
conditions, then an image is in-scope if it satisfies
all necessary conditions.
- If the concept is defined by a list of positive and
negative conditions, then an image is in-scope if it
satisfies at least one positive condition and does not
 satisfy any negative condition.
- A concept can be first defined by a list of
necessary conditions, and then each necessary
condition can be further defined by a list of positive
 and negative conditions.
</define-a-concept>
<image-type>
- in-scope: images that are likely to be in-scope for
the concept.
- ambiguous: images that are borderline in-scope for
the concept.
There are some important ambiguities that the concept
definition does not articulate clearly.
- out-of-scope: images that are likely to be out-of-
scope for the concept.
</image-type>


<step1>
Examine the concept definition and previous
descriptions, propose {num_descriptions} new
categoriesof images are in-scope but are different
from the previous descriptions.
These categories should cover significantly different
categories of images from the previous descriptions.
</step1>

<step2>
Based on your reasoning in step1, write down {
num_descriptions} new descriptions for the {image_type
} category of images.
The description should be concise, specific, and clear
. You should aim for less than 20 words for this
description.
</step2>

Write your answer in a valid XML format, adhering to
the following structure:
<reasoning>
  Write your reasoning for new categories of images in
   step 1 here.
</reasoning>
<descriptions>
  Write your description for the in-scope category in
  step 2 here.
  <!-- there should be {num_descriptions} descriptions
   in total -->
  <description></description>
  <description></description>
  <description></description>
</descriptions>

Here is the concept definition you should work on:
```

```
<definition>{definition.readable_string()}</definition
>
<previous-descriptions>
{previous_descriptions_str}
</previous-descriptions>
```

Prompts that examine ambiguities of individual images

```
<role>You are an expert linguist who is good at
brainstorming creatively.</role>
<input>
You are given the definition of a visual concept,
which serves as the guideline for human raters to
determine whether an image is within the scope of this
 concept.
You will also be given an image.
</input>
<task>
As the concept owner is still actively working on the
definition of the target concept,
there are still some ambiguities in this concept
definition.
You will help examine whether an image might highlight
 important ambiguities in the concept definition and
thus should be reviewed by the concept owner,
so that the concept owner could further improve the
definition.
</task>

<step1>
Examine the image against the definition and
determines whether the image should be classified as
in-scope or out-of-scope.
</step1>
<step2>
Now assume that the concept owner actually gives a
different classification result for this image.
Examine the image against the definition and reason
what might be the important ambiguities that the
current definition fails to capture.
As the current definition is mostly correct, you
should not completely ignore the current definition,
but instead you should focus on identifying subtle but
 important ambiguities.
</step2>
<step3>
Examine your reasoning in step2, and determine how
likely these ambiguities actually make sense.
This means that the concept owner is likely to be
unclear about his definition at this point, or this
point is likely to cause confusion to human raters.
Some images are actually clear-cut in-scope or out-of-
scope examples--in these cases, your ambiguities might
 not make much sense.
</step3>
<step4>
If you believe that the ambiguities are important in
step2, pick the most important ambiguity from step2
and summarize it in one sentence less than 30 words.
Your summary should directly point out the elements
that might cause the ambiguity in the image and the
specific requirements in the definition.
for instance,
- "The image shows two people use sign language to
communicate with each other, but it is unclear whether
 sign language is considered as "chatting"",
- "The image shows a set of cartoon dogs, but it is
unclear whether cartoon is considered as "dog" or not
."
But if you believe that the ambiguities are not
important, then your summary should be an empty string
.
</step4>
```

```
Provide your answer in the following XML structure:
<classification>The classification result of the image
 and your reasoning.</classification>
<counter-reasoning>Your reasoning about why the image
might have been misclassified at step2</counter-
reasoning>
<examination>Your reasoning about whether the
ambiguities are important at step3</examination>
<summary>Your summary of the most important ambiguity
if your answer is "yes" at step3; otherwise, leave it
empty.</summary>

Here is the definition you should work on:
<visualConceptDescription>{definition.readable_string
()}</visualConceptDescription>
```

E.4. Concept refinement

Prompts that articulate user feedback into explicit rationales

```
<role>You are an expert linguist</role>.
<input>
  You are given the definition of a visual concept and
   an image caption.
  Guided by this concept definition, human raters are
  asked to determine whether an image is within the
  scope of this concept, and write down their
  rationales and decisions.
  We also asked the concept owner to directly rate
  whether this image is in-scope or out-of-scope of
  this concept [ground-truth]
  The concept owner might also provide feedback
  regarding what the human raters should have noticed.
</input>

<task>
  The concept owner is still actively working on the
  definition of the target concept.
  The final goal is to enable human raters to
  interpret this definition to rate images exactly as
  the concept owner would do.
  Your task is to help articulate what this concept
  owner wants to clarify for this visual concept.
</task>

<define-a-concept>
Before answering the question, it is a must for you to
 understand how we define a concept in a structured
and iterative way.
- If the concept is defined by a list of necessary
conditions, then an image is in-scope if it satisfies
all necessary conditions.
- If the concept is defined by a list of positive and
negative conditions, then an image is in-scope if it
satisfies at least one positive condition and does not
 satisfy any negative condition.
- A concept can be first defined by a list of
necessary conditions, and then each necessary
condition can be further defined by a list of positive
 and negative conditions.
</define-a-concept>

<step1>
Within the context of this image, reason over what
clarifications the concept owner might want to
incorporate into the definition.
You should try to refer to specific elements in the
image to better ground your reasoning.
Your reasoning should be based on the following
questions:
1) If the concept owner provides a clear feedback,
what do you think the concept owner wants to clarify?
  Do not generalize too much beyond what the concept
  owner says.
2) If the concept owner provides a different rating
than the human raters, what is the possible reason for
 this disagreement?
  Do not generalize too much beyond this disagreement
  between ratings.
3) When the concept owner provides no clear feedback
and the human raters and the concept owner are in
agreement,
  what does this agreement between the concept owner
  and the human raters confirm?
  Especially in this scenario, since there is no less
  clear information,
  you should be more conservative and specific, and
  try to avoid generalizing too much.
</step1>
<step2>
Summarize your reasoning in the step 1 with a few
sentences.
Your summary will be used in downstream steps so make
sure it covers all the necessary information.
Please make sure that you ground your summary with
specific elements or examples in the image.
this will help others better understand your reasoning
.
</step2>

Provide your answer in a valid XML format, adhering to
 the following structure:
  <reasoning>Describe your reasoning in the step 1</
  reasoning>
  <clarification>Provide your answer to the question
  in the step 2</clarification>

<conceptDefinition>{definition.readable_string()}</
conceptDefinition>
<raterResponses>
  <decision>{ImageClassifier.rating_to_label(
  reflection_info['decision'])}</decision>
  <summary>{reflection_info['summary']}</summary>
</raterResponses>
<conceptOwner>
  <ground-truth>{ImageClassifier.rating_to_label(
  reflection_info['groundtruth'])}</ground-truth>
  <user-feedback>{reflection_info['feedback']}</user-
  feedback>
</conceptOwner>
<image_caption>{image.image_caption}</image_caption>
```

Prompts that generate refinement candidates

```
<role>You are an expert linguist</role>.
<input>
 You are given the definition of a visual concept,
 which serves as the guideline for human raters to
 determine whether an image is within the scope of
 this concept.
 However, as the concept owner is still actively
 working on the definition of the target concept,
 human raters incorrectly rated an image regarding a
 particular focus concept within this definition.
 Therefore, the concept owner wants to clarify their
 points and improve the definition of this focus
 concept.
 They provide their clarifications for each of {
 images_num} images respectively.
 The final goal is to have a more accurate and
 comprehensive concept definition so that human
 raters can rate images exactly as the concept owner
 would do.
</input>

<task>
 Your task is to determine how to improve the
 definition of the focus concept in the most accurate
  and concise way.
</task>

<define-a-concept>
Before answering the question, it is a must for you to
 understand how we define a concept in a structured
and iterative way.
- If the concept is defined by a list of necessary
conditions, then an image is in-scope if it satisfies
all necessary conditions.
- If the concept is defined by a list of positive and
negative conditions, then an image is in-scope if it
satisfies at least one positive condition and does not
 satisfy any negative condition.
- A concept can be first defined by a list of
necessary conditions, and then each necessary
condition can be further defined by a list of positive
 and negative conditions.
</define-a-concept>

<step1>
Examine the clarifications for all these images and
summarize the key points that the concept owner might
want to incorporate into the definition.
Some clarifications might be similar, so you should
aggregate them;
On the other hand, some clarifications might be very
different, so you should consider listing them
separately.
</step1>
<step2>
Based on your answers in previous steps, reason how to
 incorporate the key points into the definition.
You could either choose to add a new positive or
negative signal to the definition, or modify an
existing positive or negative signal.
For an existing positive or negative signal, you
should not add a child positive or negative signal to
it.
For an existing necessary condition, you should not
add a sibling positive/negative/necessary signal to it
.
</step2>
<step3>
Based on your answers in previous steps, write down
the changes you want to make to the definition.

You should be careful about the language of your
changes, in particular, there are several requirements
.
<description-requirements>
 1) Always make sure that your final description is
 CONCISE, COHERENT, and ACCURATE; an average person
 could easiy determine whether an image satisfies the
  signal based on the description.
 2) DO NOT write a complex sentence structure in a
 description of a signal.
 3) You should only make important changes to the
 description.

 If the original description misses a point, you are
 encouraged to use one of the following ways to
 incorporate the nuances the concept owner wants to
 convey:
 a) add new adjectives, b) use different verbs, or c)
  add a few constraint words.
 If the original description uses an ambiguous or
 misleading word or example, you are encouraged to a)
  refine the word or example, or b) simply remove
 them from the description.

 4) Be careful about your word choices of verbs,
 nouns, or adjectives, which might carry unexpected
 nuances.
    e.g., be careful about using 'depict' or 'mention
    ', or 'show' as the previous two verbs introduce
    the slight emphasis on visual or textual aspects.
    e.g., be careful about using adjectives like '
    clearly' or 'explicitly' as they might suggest a
    degree of visibility to the original concept.
 5) If your description consists of two independent
 conditions, you might consider use the format like "
 Images that 1) ... and 2) ..." to make it more clear
 .
</description-requirements>

</step3>

Provide your answer in a valid XML format, adhering to
 the following structure:
 <keypoints>Describe your reasoning of the key points
  of these clarifications in the step1</keypoints>
 <reasoning>Describe your reasoning of how to
 incorporate the key points into the definition in
 the step2</reasoning>
 <improve-description>
  The improved description of the visual concept in
  the step3.
  You should only write down changes you proposed in
  the following format.
  1) If you want to edit an existing signal, the
  format is as follows:
  <concept>
   <name>The name of the signal you want to edit</
   old-name>
   <old-description>The original description of the
   signal</old-description>
   <new-description>The new description of the
   signal</new-description>
  </concept>
  2) If you want to add a new signal, the format is
  as follows:
  <concept>
   <parent-signal>The name of the parent signal</
   parent-signal>
   <type>The type of the new signal, either '
   positive' or 'negative'</type>
   <new-name>The new name of the signal</name>
   <new-description>The new description of the
   signal</description>
  </concept>
  It might be possible that you need to make multiple
   changes, so you should write down all of them.
 </improve-description>

<conceptDefinition>{definition.print_definition()}</
conceptDefinition>
<clarifications>{reflections_str}</clarifications>
```

F. System Parameters

Borderline Image Retrieval Module. To select which cluster to query next, we compute an interaction-based score for every cluster using four heuristics derived from users' previous interactions. For each cluster, we track how many images have been explored, how often users marked them as mistakes, how often users provided textual feedback, and the distribution of user-provided ratings. We then compute:

- **Mistake rate.** The proportion of explored images that users labeled as incorrect for the current definition. Clusters with higher mistake rates are prioritized because they reveal boundary mismatches.
- **Feedback rate.** The proportion of explored images for which users provided any written justification. Clusters that elicit richer feedback are treated as more informative.
- **Exploration value.** A term that favors clusters with many unseen images. This prevents repeatedly sampling from clusters that are already heavily explored.
- **Diversity rate.** The standard deviation of user-provided ratings within the cluster. Higher diversity indicates conceptual ambiguity and thus higher expected value for refinement.

These four components are combined into a weighted score:

$$\begin{aligned}\text{Score} = {} & 0.5\,\text{mistake\_rate} + 0.3\,\text{feedback\_rate} \\ & + 0.15\,\text{exploration\_value} \\ & + 0.05\,\text{diversity\_rate}.\end{aligned} \tag{12}$$

where the coefficients were chosen based on preliminary experiments balancing exploration and refinement. The cluster with the highest score is selected for the next round of iteration.

Ambiguity mining. To surface a coherent batch of 5 borderline images per round, we sample 25 candidates, generate short ambiguity summaries, embed them, and cluster the embeddings to isolate a single ambiguity dimension. We use an adaptive DBSCAN radius to identify a cluster of at least 5 images. The full clustering procedure is provided below.

Concept refinement. Based on the user's feedback for the current batch, the system generates 5 candidate refinements $\{d_t^{(1)}, \ldots, d_t^{(5)}\}$ of the current definition d_t. We first filter the top 3 candidates based on their performance on the current borderline set $\mathcal{B}_t$, ensuring that the new rationales are directly reflected. Among these 3 candidates, we then select the final update d_{t+1} according to their performance on the full labeled set $\mathcal{L}_t$ accumulated so far. This two-stage procedure incorporates user feedback immediately while preserving global consistency across rounds.

Algorithm 1 Adaptive DBSCAN for Ambiguity Clustering

Require: embeddings $\{e_i\}$; minimum cluster size $k = 5$; minimum radius $\varepsilon_{\min} = 0.2$; maximum radius $\varepsilon_{\max} = 0.8$; step size $\Delta\varepsilon = 0.01$
Ensure: a cluster of size $\geq k$

```
ε ← ε_min
while ε ≤ ε_max do
    labels ← DBSCAN(ε) on {e_i}
    clusters ← non-noise groups in labels
    if any cluster has size ≥ k then
        return that cluster
    end if
    ε ← ε + Δε
end while
return clusters from final ε
```

G. Example Definitions

Here is one full example of the definition d_0 after the scoping stage.

This includes any of the following visual elements:

- *Healthy Dish*: Images show a prepared meal or dish that is prominently composed of healthy ingredients, such as salads, grain bowls, grilled or steamed lean proteins (e.g., chicken breast, fish, or tofu) with vegetables, or oatmeal with fruit.
- *Healthy Beverages*: Images show healthy beverages, such as smoothies made from fruits and vegetables, freshly squeezed juices, or fruit-infused water.

However, the following visual elements are excluded:

- *Processed Food*: Images show processed foods typically considered unhealthy, such as pizza, mayonnaise, or whipped cream.
- *Raw Ingredients*: Images show farming scenes, raw ingredients, or meat that is not yet prepared as edible food.
- *Not Focus on Food*: Images show an activity related to food where the food itself is not the main subject, such as people cooking, shopping in a grocery aisle, or dining in a restaurant.