A Zero-Shot Learning Approach for Ephemeral Gully Detection from Remote Sensing using Vision Language Models

Seyed Mohamad Ali Tousi[†*], Ramy Farag[†*], Jacket Demby's[†], Gbenga Omotara[†],
John A. Lory[‡], G. N. DeSouza[†]

[†]Vision Guided and Intelligent Robotics Laboratory (ViGIR), EECS Dept.

[‡]Division of Plant Sciences

University of Missouri Columbia - MO - US

stousi, rmf3mc, udembys, goowdf, loryj, desouzag@missouri.edu

Abstract

Ephemeral gullies are a primary cause of soil erosion and their reliable, accurate, and early detection will facilitate significant improvements in the sustainability of global agricultural systems. In our view, prior research has not successfully addressed automated detection of ephemeral gullies from remotely sensed images, so for the first time, we present and evaluate three successful pipelines for ephemeral gully detection. Our pipelines utilize remotely sensed images, acquired from specific agricultural areas over a period of time. The pipelines were tested with various choices of Visual Language Models (VLMs), and they classified the images based on the presence of ephemeral gullies with accuracy higher than 70% and a F1-score close to 80% for positive gully detection. Additionally, we developed the first public dataset for ephemeral gully detection, labeled by a team of soil- and plant-science experts. To evaluate the proposed pipelines, we employed a variety of zero-shot classification methods based on State-of-the-Art (SOTA) open-source Vision-Language Models (VLMs). In addition to that, we compare the same pipelines with a transfer learning approach. Extensive experiments were conducted to validate the detection pipelines and to analyze the impact of hyperparameter changes in their performance. The experimental results demonstrate that the proposed zero-shot classification pipelines are highly effective in detecting ephemeral gullies in a scenario where classification datasets are scarce.

1. Introduction

Soil erosion is a geomorphological land degradation process that can result in environmental harm, property damage, loss of livelihoods and services, and social and economic challenges. Ephemeral gully erosion reduces the sustainability of agricultural systems and causes significant sediment-related issues downstream [22]. Ephemeral gullies remain a prominent and hard-to-treat cause of soil erosion [14]. A fast, reliable and cost-effective method to detect ephemeral gullies in the agricultural fields could play a crucial role in preventing the soil erosion and successful execution of soil conservation practices. A detection system using remote sensing images to distinguish between agricultural fields that show signs of ephemeral gully formation versus ephemeral gully-free agricultural fields will advance efforts of soil researchers, farmers and responsible agencies to study, treat and/or prevent soil erosion.

The emergence of Foundation Models, Vision Language Models (VLM), and Visual-Question-Answering (VQA) has great potential for tackling a wide range of computer vision challenges, particularly in remote sensing ([8, 20, 35]). When classifying nuanced and subjective concepts, Foundation Models and VLMs tend to hallucinate. To address this problem, techniques such as Modeling Collaborator (MC) [31] have been introduced. One of the advantages of MC is that it does not require an extensive amount of labeled data which is especially attractive to us since labeled datasets for ephemeral gully detection are extremely scarce. The lack of sufficient datasets for detecting ephemeral gullies to train conventional computer vision models leads to the exploration of methods that do not rely heavily on extensive labeled data. Techniques such as zero-shot classification, and transfer learning are common alternatives.

We assessed SOTA open-source zero-shot, transfer learning, and VLM-based classification methods in the task of classifying remote sensing images with signs of ephemeral gullies in agricultural fields. Additionally, we are providing the first dataset for detecting ephemeral gullies using remote sensing images, labeled by a group of soil

[*]These authors contribute equally to the paper.

and plant science experts.

Our key contributions are as follows:

1. We present and evaluate three classification pipelines for detecting ephemeral gullies in remote sensing images, exploring zero-shot, transfer learning, and Vision-Language Models (VLMs). To our knowledge, this is the first work to successfully address this under-explored task. Inspired by [31], we further integrate large language models (LLMs) to enhance reasoning in the classification process, using VLMs for Visual Question Answering (VQA) and LLMs for reasoning, and compare their performance against existing zero-shot classification baselines.

2. We conduct extensive experiments to investigate the choices of VLMs, types and number of questions, and aggregation methods for VQAs. Those experiments lead to a better understanding of the advantages and limitations of zero-shot classification methods such as [31].

3. We make available the first public dataset for the detection of ephemeral gullies using remote sensing images. The dataset comprises high-resolution remote sensing RGB images of agricultural locations over different periods of time. The images are labeled by a team of soil and plant science experts to establish the ground truth classifications.

2. Background and Related Work

In this section, we will explore the existing research articles in the literature on ephemeral gully detection (assuming its presence in the image), and ephemeral gully formation assessments using remote sensing imagery. We also explore the SOTA zero-shot classification methods and VLMs that form the foundation of this research.

2.1. Ephemeral Gullies in Remote Sensing

Remote detection of ephemeral gully through classification remains largely unexplored, though manual assessment and evaluation are well-documented in agricultural literature [6,9,13,16]. However, the time commitment associated with manual assessments limits their scalability for large-scale applications. Authors in [25] tested the effectiveness of using Google Earth temporal images to assess ephemeral gullies in U.S. agricultural fields, presenting a systematic method for identifying and locating these gullies within a certain error margin. Their approach relied heavily on manual labor and utilizes Google Earth images, which lack the image standardization needed for remote sensing in computer vision contexts.

Some researchers have attempted to automate parts of the process. There is extensive research on ephemeral gully risk assessment using topographic index models to identify areas of the field most susceptible to formation of ephemeral gullies [28]. There are examples of computer vision techniques applied to detect ephemeral gullies. For example, [36] employed conventional computer vision techniques, such as directional edge detection, to create a ”semiautomatic” method for identifying ephemeral gullies. Manual pre-identification of areas containing gullies was necessary due to the faint visibility of these features. Similarly, [15] applied deep learning methods like Convolutional Neural Networks (CNNs) [10] and U-Nets [26] for delineating gullies in non-agricultural areas. Current model- and image-based methods require either post-identification verification of presence of a gully or manual pre-identification of gully-prone regions. To date, the core task of ephemeral gully classification continues to rely on human intervention.

2.2. VLMs and Zero-Shot Classification

Authors in [17, 18] introduced Visual Instruction Tuning through their LLaVA (Large Language and Vision Assistant) model, aimed at developing a general-purpose vision-language assistant. LLaVA integrates a CLIP-based visual encoder with a language model, specifically Vicuna [4, 21], fine-tuned using multimodal instruction-following data generated by GPT-4 [1]. This combination enables the model to effectively process and understand both visual and textual inputs, enhancing performance across diverse vision-language tasks.

Zero-shot classification allows deep learning models to classify instances without prior training on the same distribution of instances. Significant advancements in this field have been achieved through models like CLIP (Contrastive Language–Image Pre-training) [24], which employs vast internet-sourced image-text pairs to align image and text embeddings in a unified multi-modal space. This alignment enables robust zero- and few-shot transfer performance for various vision tasks without fine-tuning. Extensions of CLIP [12, 23, 30] continue to refine this alignment for improved downstream task performance. Building on this foundation, [27] introduced AdaptCLIPZS, a framework that enhances zero-shot classification in fine-grained domains by incorporating detailed, category-specific descriptions generated by LLMs. These descriptions are integrated into context-rich prompts during VLM fine-tuning, consistently improving performance across benchmarks.

Modeling Collaborator (MC) [31] has demonstrated strong results in nuanced and subjective zero-shot classification tasks. Combining Visual-Question-Answering (VQA) with LLMs, MC avoids common large-model hallucinations in subjective tasks by using a series of Yes/No questions derived from the classification concept. These responses are aggregated by an LLM to produce labels that can be directly used or employed in training end classifiers.

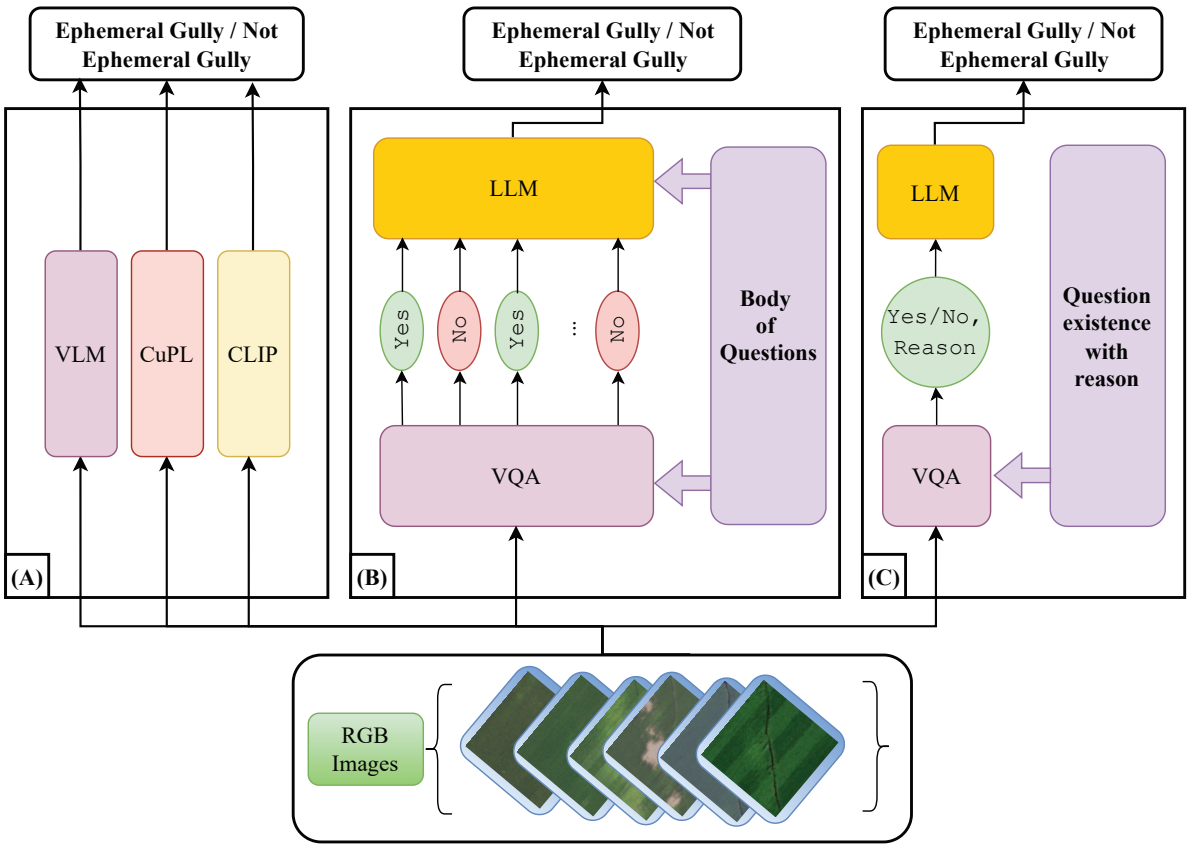


Figure 1. The proposed independent pipelines to detect ephemeral gullies. We are exploring three types of Pipelines; **(A)**: Feeding the remote sensing RGB images directly to SOTA zero-shot classification methods such as CLIP [24], CuPL [23], and a variety of VLMs including Llama 3.2-Vision [32], Qwen [3], and Llava [18]. **(B)**: Inspired by [31], we pass the RGB images through a VQA system, which responds to a series of Yes/No questions regarding visual attributes indicative of ephemeral gullies. An LLM then aggregates these responses to classify the images. **(C)**: Similar to (B), we propose asking the VLM a single descriptive question about the image, with an LLM interpreting the response to make a final classification.

Unlike manual-label-heavy approaches such as Agile Modeling (AM) [29], MC achieves comparable results with only 100 labeled samples for validation. This framework excels at classifying subjective visual concepts, outperforming traditional zero-shot and AM methods in both accuracy and efficiency.

The rest of the paper is organized as follows; In section 3 we will present the proposed pipelines to detect ephemeral gully. The experimental setup, implementation details and experimental results will be presented in section 4. Then, in sections 5 and 6 we will present a discussion about the experimental results and conclusion of the paper respectively.

3. Proposed Method

Given a set of RGB images $I = \{i_1, i_2, ..., i_M\}$ representing M temporal images (tile) at one specific location, we create a dataset $\mathcal{D} = \{(I_n, y_n)|n \in \{1, 2, ..., N\}\}$ consisting of N locations, in which each location has a true label y_n, determined by a group of soil and plant science experts looking at I_n. Figure 1 shows the three classification pipelines proposed and evaluated in this work, which are presented in the next three sections.

3.1. SOTA Zero-Shot Classification Methods

In the first classification pipeline that acts as the baseline for our application (pipeline **A** in Figure 1), we will use a variety of zero-shot classification methods, such as CLIP [24], CuPL [23] and multiple VLMs like Llama 3.2 - Vision [32], Qwen [3], and Llava [18])

The remote sensing RGB images are submitted to those models to produce a label $\hat{y}_n$ for each location I_n. Equation 1 describes this pipeline.

$$\hat{y}_n = VLM(I_n, p) \tag{1}$$

Where p is the input prompt to the VLM for detecting ephemeral gullies [1].

[1]CLIP [24] requires a positive and a negative prompt to produce the label for each I_n. Other VLMs can be prompted directly to produce the

3.2. Visual Attributes + Reasoning

Building on [31], the second pipeline proposed and evaluated in this research is designed as an end-to-end process (see pipeline **B** in Figure 1). This pipeline consists of two key components: 1) *Visual Attributes Understanding* and 2) *Reasoning*. The initial step in this pipeline involves identifying distinctive visual attributes of ephemeral gullies to generate a *body of questions* centered around these attributes. These questions help mitigate hallucinations when using baseline VLMs. The VLM processes the images alongside the generated questions, providing answers based on the image content. In the Reasoning component, these answers are aggregated by an LLM to determine the predicted class.

3.2.1 Visual Attributes Understanding

The goal of the first component of pipeline **B** is to:

1. Identify specific visual attributes of a formed ephemeral gully.
2. Create a *Body of Questions*, $Q = (q_1, q_2, ..., q_K)$ consisting of K Yes/No questions based on the identified visual attributes.
3. Use the VLM to generate answers to Q based on I_n.

In [23], the authors proposed to use an LLM (e.g. GPT-4 [1]) to produce the specific visual attributes associated with each class. In this work, we use a combination of attributes generated by GPT-4 and visual aspects coming from plant and soil science experts. Once we obtain a set of visually distinguishable attributes for the ephemeral gully class, we build a body of questions in which we have one or more questions targeting each of the attributes.

Once Q is obtained, answers are generated by a VLM:

$$\boldsymbol{A}_n = VLM(I_n, Q) \tag{2}$$

Where $\boldsymbol{A}_n$ is the set of answers for all of the Yes/No questions in Q, based on I_n contents. Subsequently, we can proceed to do reasoning based on given answers to our questions.

3.2.2 Reasoning

In the reasoning component of pipeline **B**, an LLM model aggregates the answers of Q and produce the final prediction:

$$\hat{y}_n = LLM(\boldsymbol{A}_n, Q, p) \tag{3}$$

Where p is the aggregation prompt, and $\hat{y}_n$ is the predicted class for I_n.

label.

3.3. A Descriptive Question + Reasoning

The last pipeline investigated in this research involves formulating a descriptive question to be submitted to the VLM for reasoning over the content of image location I_n (refer to pipeline **C** in Figure 1). Subsequently, an LLM evaluates the provided reasoning to determine its validity and uses it to produce the final label $\hat{y}_n$.

3.4. Transfer Learning Using Visual Attributes

For comparison with the previous three proposed pipelines, we devised a non-zero-shot approach based on transfer learning. The rationale for this fourth approach was to contrast zero-shot and non-zero-shot methods with the intent of evaluating the performance of zero-shot methods. This approach uses the answers to questions Q generated by the VLM as extracted features from the contents of image location I_n. These features are then used as inputs to train a simple Multi Layer Perceptron (MLP) on the development (Dev) set to produce the final label $\hat{y}_n$. The results of this approach are compared with the zero-shot methods discussed earlier, demonstrating that the zero-shot algorithms produce similar performance to this learning-based approach.

4. Experimental Setup and Results

As mentioned earlier, the main contribution of our work are: the use of VLMs as zero-shot classification approaches for detecting ephemeral gullies; exploring different types of queries using VQA, and an extensive labeled dataset. In order to evaluate the efficacy of the proposed pipelines and the queries used for VQA, we must address the following questions:

1. How well each pipeline performs the task of ephemeral gully detection?
2. What are the impacts of specific choices of VLMs and LLMs in those pipelines?
3. How the types and number of questions affect the performance of the pipelines?
4. How does the performance of the proposed pipelines compare to a non-zero-shot approach based on transfer learning?

To answer these questions, we conducted the following experiments to establish the validity and reliability of the classification pipeline.

4.1. Evaluation Dataset

We used USDA National Agriculture Imagery Program (NAIP) [33] to gather 6 temporal high-resolution (1m) RGB

Parameter	Dev Set	Test Set
Total Number of Locations	310	311
Number of EG-Positive Locations	177	177
Number of EG-Negative Locations	133	134
Number of Images per Location (M)	6	
Image Size	128x128	
Image Spatial Resolution	1m	

Table 1. The Evaluation Dataset characteristics for Dev and Test sets.

images for agricultural areas over a period of 10 years. Google Earth Engine [7] was used to access and download the dataset. Once the dataset was obtained, a team of 8 soil and plant science experts labeled each of the areas that showed a sign of ephemeral gully formation. Then, we divided the acquired dataset into two separate Dev and Test sets. Table 1 shows the statistics of the final evaluation dataset, after the labeling process and pruning the outliers. The dataset is available to download through this link: http://vigir.missouri.edu/Research/GullyDetection/.

4.2. Implementation Details

This section provides the implementation details used to answer the research questions, as well as the experimental setups.

4.2.1 Processing the RGB Images Simultaneously

The primary implementation challenge to address was enabling VLMs to process six images as their inputs simultaneously. Since identifying ephemeral gullies relies on temporal information, processing each RGB image independently and aggregating the results is insufficient[2]. Moreover, most of the open-source SOTA VLMs used in this study do not natively support multiple images at the time of writing. There are at least two potential solutions for this problem; The first solution involves encoding all six images into the embedding space, concatenating the embeddings, and then supplying the concatenated embedding vector as the conversation context when prompting the VLMs. However, this approach is constrained by the limited context length supported by VLMs—at most 4096 tokens at the time of writing—which makes it impractical to embed multiple separate RGB images.

The second solution is to create a collage of the images and pass it to the VLM as a single image. Our experiments prove this approach effective when working with various VLMs, including CLIP [24], Llama 3.2-Vision [32], and Qwen [3].

[2]This is the default way that CLIP [24] uses to process multiple images.

4.2.2 Body of Questions

As mentioned in section 3.2.1, we are using a combination of GPT-4 [1] and soil and plant science experts to produce the body of questions (Q). The final Q is as follows:

Body of Questions (Q)

Given these six images of the exact same area and collected over a period of 10 years ...

1. Do you see a low point in the terrain?
2. Do narrow, winding paths or channels appear?
3. Are there winding paths that become intermittent recurrent?
4. Are there any linear depressions or ruts which appear more pronounced along natural drainage lines or slopes?
5. Are there narrow and shallow channels which appear intermittently deeper or more indented into the soil?
6. Are there areas where soil appears disturbed or vegetation is removed?
7. Does a specific path lack vegetation, suggesting an evolving or emerging channel?
8. Are there varying types and levels of coarseness in the texture of the soil?
9. Are there clear starting and ending points of potential channels?
10. Are there small rills or grooves indicating water flow?
11. Is there a varying exposure of lighter or darker colored soil?
12. Are there sediment accumulations forming?
13. Are there signs of water activity, like soil clumps or crusting, that appear or intensify in specific areas?
14. Are there any branching patterns that resemble temporary streams?
15. Are there indications of nearby human activity, such as tillage or machinery tracks?
16. Do you see any sign of water flow patterns across the field in multiple images?
17. Do you see any edges in the images indicating removal of soil along the water pathway?
18. Do you see any cuts in the soil associated with water flow across the field?
19. Do you see any indication of human activity such as tillage that is not naturally happened in the field?

Parameter	Value
Number of Temporal Images (M)	6
Number of Questions (Q size)	19
VLM Choices	Llama 3.2 - Vision [32] Qwen [3] Llava [18]
LLM Choices	Llama 3.2 [32] Qwen [3]

Table 2. The experimental setup and model choices for exploring classification pipelines.

4.2.3 Experimental Setups and Model Choices

We utilize LLava [11], Llama 3.2-Vision [32], and Qwen [3] as our VLMs. These models were selected primarily based on their performance on open benchmarks. At the time of writing, Llama 3.2 and Qwen rank among the top models across various vision and language benchmarks. LLava was included to facilitate a comparison between SOTA models and non-SOTA models.

We used Ollama [19] and Hugging Face [34] as our frameworks to interact with the open-source VLMs and LLMs. Table 2 shows the hyperparameters and model choices for the explored methods.

To enable the VLMs to classify the presence of ephemeral gullies within a specific agricultural area, we used two different prompts for pipelines **A**, and **C**:

Prompt for Baseline Results (Pipeline A)

```
Given this collage of six images of the
exact same area and collected over a period
of 10 years.  Are there any ephemeral gully
appearances by looking at all of them together?
Reason and conclude with only yes or no.
```

Prompt for Single Question Reasoning (Pipeline C)

```
Given this collage of six images of the exact
same area and collected over a period of
10 years.  Are there any ephemeral gully
appearances by looking at all of them
together?.  Provide the reasons for your
answer.
```

4.3. Experimental Results

To evaluate the performance of the classification pipelines and to address research questions **(1)** and **(2)**, we conducted experiments measuring the accuracy of each pipeline on the test set and analyzed the impact of different model choices for VLMs and LLMs. Additionally, to address question **(4)**, we trained a simple MLP on the Dev set to aggregate answers from the VQA step, with results presented in Table 3. Given the dataset's imbalance in positive and negative images, relying solely on the F1 score could be misleading; therefore, we included the Macro F1 score in Table 3 to reflect a class-averaged evaluation.

To address research question **(3)**, we performed two additional experiments. First, an expert ranked questions in Q by their relevance for distinguishing ephemeral gullies, creating subsets of 3, 6, 9, 12, 15, and 18 questions, with their indices listed in Table 4. Second, a hyperparameter optimization framework (Optuna [2]) was used to identify the optimal set of questions for improving pipeline **B**. Performance results for these experiments are summarized in Table 5. To further analyze the decision-making process, a histogram in Figure 2 illustrates the frequency of "Yes" responses for each question across positive and negative test and Dev sets.

5. Discussion

5.1. Performance of The Classification Pipelines

Both CLIP and CuPL presented difficulty detecting ephemeral gullies (Table 3). This was expected given the limited representation of remote sensing images with delineated ephemeral gullies in their pre-training datasets. CuPL, which builds on CLIP, inherited the same challenges in distinguishing ephemeral gullies. Similarly, Llava and Llama 3.2-Vision performed poorly in pipeline **A** due to biases in their classification behavior. Llava classified most areas as negatives, yielding a low F1 score for the EG-positive class, while Llama 3.2-Vision labeled most locations as EG-positive, reducing the F1 score for the EG-negative class. Both models exhibited low Macro F1 scores, highlighting the importance of pre-training datasets in zero-shot classification methods, as these datasets lacked sufficient information about ephemeral gullies' visual characteristics.

In contrast, Qwen2-VL demonstrated strong performance in pipeline **A**, suggesting that its pre-training dataset contained richer information about ephemeral gullies. We have also achieved promising results in pipeline **B**, particularly when using Llama 3.2-Vision as the VLM and Llama 3.2 as the LLM, performing comparably to Qwen2 in pipeline **A**. However, substituting either the VLM or LLM with Qwen2 reduced performance, as discussed in Section 5.3. Pipeline **C** did not surpass the performance of pipeline **A**, indicating that prompting VLMs to reason about their decisions may exacerbate hallucination rather than mitigate it. Additionally, a transfer learning-based method failed to outperform zero-shot classifiers in pipeline **A** (with Qwen2-VL) and pipeline **B** (with Llama 3.2 - Vision), suggesting that zero-shot methods are particularly effective for detecting ephemeral gullies in data-limited scenarios.

P#	VLM	LLM	Experiment	TP	FP	FN	TN	Prec.	Rec.	Acc.	F1 (G)	F1 (NG)	Macro F1
			Input = Collage of 6 Temporal Aerial Images										
(A)	CuPL	-	CuPL (GPT4-Prompts)	9	4	168	130	0.692	0.051	0.447	0.095	0.602	0.348
(A)	CLIP	-	CLIP (Pos/Neg)	134	90	43	44	0.598	0.757	0.572	0.668	0.398	0.533
(A)	Llava-Llama3-8b	-	VQA (1Y/N - Q&A)	22	20	155	114	0.524	0.124	0.437	0.201	0.566	0.383
(A)	Llama3.2-90b	-	VQA (1Y/N - Q&A)	146	103	31	31	0.586	0.825	0.569	0.685	0.316	0.501
(A)	Qwen2-VL-72B	-	VQA (1Y/N - Q&A)	167	79	10	55	0.679	0.944	**0.714**	**0.790**	0.553	**0.671**
(B)	Llava-Llama3-8b	Llama3.2	VQA(15Y/N - Q&A)	177	134	0	0	0.569	1.000	0.569	0.725	0.000	0.363
(B)	Qwen2-VL-72B	Llama3.2	VQA(15Y/N - Q&A)	173	103	4	31	0.627	0.977	0.656	0.764	0.367	0.565
(B)	Llama3.2-Vision	Qwen2	VQA(15Y/N - Q&A)	25	7	152	127	0.781	0.141	0.489	0.239	0.615	0.427
(B)	Qwen2-VL-72B	Llama3.2	VQA(15Y/N - Q&A)	177	134	0	0	0.569	1.000	0.569	0.725	0.000	0.363
(B)	Llama3.2-90b	Llama3.2	VQA(15Y/N - Q&A)	110	44	67	90	0.714	0.621	0.643	0.665	**0.619**	0.642
(C)	Llava-Llama3-8b	Llama3.2	VQA (1Q + Reason)	177	134	0	0	0.569	1.000	0.569	0.725	0.000	0.363
(C)	Llama3.2-90b	Llama3.2	VQA (1Q + Reason)	130	101	47	33	0.563	0.734	0.524	0.637	0.308	0.473
(TL)	Llama3.2-90b	-	VQA(15Y/N) + MLP	117	50	60	84	0.701	0.661	0.646	0.680	0.604	0.642

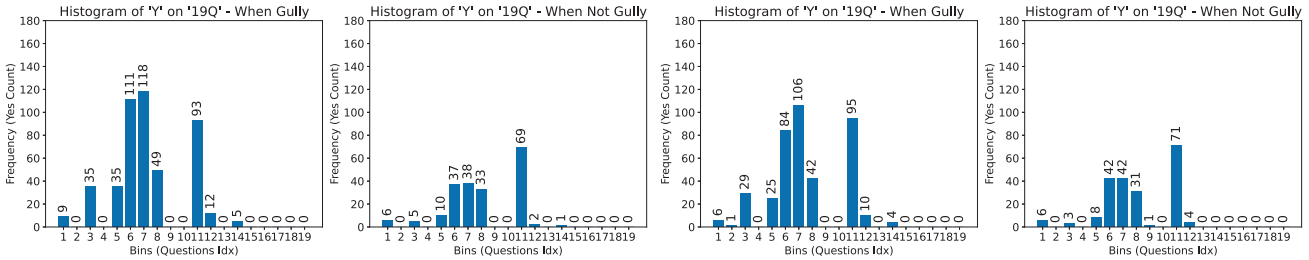
Table 3. Performance of VLMs on the Ephemeral Gully classification dataset. P# column indicates the pipelines as designated in Figure 1. In each pipeline, the best metric scores are underlined. The best metric scores across all the pipelines are in bold. (TL) stands for the transfer learning method which is training an MLP on the Dev set to aggregate the VQA results.

Experiment	Question Indices
VQA(3Y/N - Q&A + LLM)	2, 3, 14
VQA(6Y/N - Q&A + LLM)	2, 3, 14, 5, 4, 6
VQA(9Y/N - Q&A + LLM)	2, 3, 14, 5, 4, 6, 9, 11, 13
VQA(12Y/N - Q&A + LLM)	2, 3, 14, 5, 4, 6, 9, 11, 13, 8, 10, 7
VQA(15Y/N - Q&A + LLM)	2, 3, 14, 5, 4, 6, 9, 11, 13, 8, 10, 7, 1, 12, 15
VQA(18Y/N - Q&A + LLM)	2, 3, 14, 5, 4, 6, 9, 11, 13, 8, 10, 7, 1, 12, 15, 16, 17, 18

Table 4. Partition of the 19 Questions per index for the experiments in Table 5.

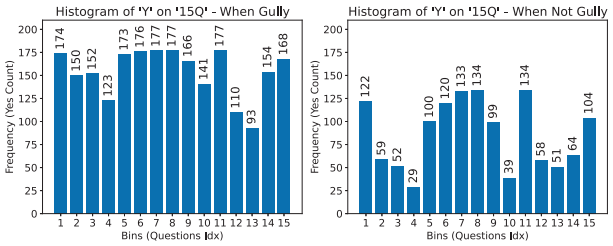
Model	Experiment	TP	FP	FN	TN	Prec.	Rec.	Acc.	F1 (G)	F1 (NG)	Macro F1
	Collage of 6 Temporal Aerial Images										
Ours (Llama3.2-90b)	VQA(3Y/N - Q&A + LLM)	29	3	148	131	0.906	0.164	0.514	0.278	0.634	0.456
Ours (Llama3.2-90b)	VQA(6Y/N - Q&A + LLM)	40	9	137	125	0.816	0.226	0.531	0.354	0.631	0.493
Ours (Llama3.2-90b)	VQA(9Y/N - Q&A + LLM)	37	9	140	125	0.804	0.209	0.521	0.332	0.627	0.479
Ours (Llama3.2-90b)	VQA(12Y/N - Q&A + LLM)	111	44	66	90	0.716	0.627	0.646	0.669	0.621	0.645
Ours (Llama3.2-90b)	VQA(15Y/N - Q&A + LLM)	110	44	67	90	0.714	0.621	0.643	0.665	0.619	0.642
Ours (Llama3.2-90b)	VQA(18Y/N - Q&A + LLM)	37	10	140	124	0.787	0.209	0.518	0.330	0.623	0.477
Optuna (Llama3.2-90b) $\Rightarrow$ **4Y/N: (3,6,7,12)**	VQA(4Y/N - Q&A + LLM)	110	44	67	90	0.714	0.621	0.643	0.665	0.619	0.642

Table 5. Performance of VLMs on the Ephemeral Gully classification dataset when varying the number of questions.



(a) Yes - Gully (Valid Set) (b) Yes - Not Gully (Valid Set) (c) Yes - Gully (Test Set) (d) Yes - Not Gully (Test Set)

Figure 2. Analysis of responses across our body of 19 questions obtained with Llama3.2-90b VLM on the test set.



(a) Yes - Gully (Test Set) (b) Yes - Not Gully (Test Set)

Figure 3. Analysis of responses across our body of 15 questions obtained with Qwen2-VL-72b VLM on the test set.

5.2. Importance of Questions in Pipeline (B)

Figure 2 illustrates the frequency of "Yes" responses for each question across EG-positive and EG-negative locations. The histogram reveals that not all questions significantly impact the decision-making process, which aligned with our expectations. This is because the questions, generated using expert input and GPT-4 knowledge, do not always correspond to visual features that VLMs can reliably interpret. A subset of the questions likely contributes more meaningfully to the results, leading to a series of experiments with different subsets of questions (Table 5). Performance improved as the number of questions increased, introducing critical visual attributes—up to a certain point. Beyond 12 questions, performance began to plateau, with a slight drop observed from 12 to 15 questions, and a significant decline from 15 to 18 questions.

This behavior can be attributed to two factors. First, as the number of questions grows, redundancy becomes more likely, leading to inconsistent answers that confuse the LLM during aggregation. Second, the increased number of questions raises the potential for conflicts among them, complicating the decision-making process. To address this, Optuna was used to identify the most influential questions. Remarkably, using just the four key questions identified by Optuna produced the same results as using 15 questions, demonstrating the effectiveness of optimizing the question set for better performance (Table 5). This is a clear advantage of our pipeline over other VQA-based classification approaches like MC [31].

5.3. Qwen2-VL vs. Llama3.2-Vision

Both Qwen2-VL and Llama 3.2-Vision showed strong performance in ephemeral gully detection tasks, but their strengths varied depending on the part of the pipeline they were used in (Table 3). Qwen2-VL excelled in pipeline **A**, demonstrating a robust ability to directly detect ephemeral gullies. Its superior visual understanding aligned with benchmarks [5], which consistently rank Qwen2-VL above other open-source VLMs in visual comprehension. However, its language abilities lagged behind those of Llama 3.2, which makes the latter more effective in tasks requiring nuanced language reasoning.

Llama 3.2's linguistic proficiency enhanced its performance as an aggregator LLM, as shown in Table 3. Its superior language comprehension enabled it to generate better answers to the body of questions, making it a more effective VQA system. Conversely, Qwen2-VL's answer histograms (Figure 3) revealed a tendency to respond "Yes" to most questions, especially the key questions identified by Optuna [2], regardless of context. This overgeneralization limited its ability to distinguish between EG-positive and EG-negative tiles, often leading to the incorrect labeling of all tiles as positive.

5.4. Practical Implications and Future Works

The proposed pipelines demonstrated reliable and robust results; however, challenges remain that need to be addressed for their successful deployment in agricultural applications and real-world scenarios. In practical contexts, false negative classifications pose a greater concern than false positives, as missing ephemeral gullies in agricultural fields can lead to significant issues. To enhance the effectiveness of the proposed pipelines, improving their sensitivity (F1 (G)) to the presence of ephemeral gullies is crucial. Future research could explore the fine-tuning of Vision-Language Models (VLMs) to provide better and more robust vision and language understanding in this specific context, thereby improving the pipeline's overall performance and reliability.

6. Conclusion

This study introduced the first stand-alone pipeline to detect ephemeral gullies and an evaluation dataset labeled by experts. Our results demonstrate that VLM-based zero-shot classification methods, particularly those using Qwen2-VL and Llama 3.2 - Vision, were effective in detecting ephemeral gullies. While Qwen2-VL excelled in visual understanding, Llama 3.2 - Vision's language abilities made it a stronger aggregator in question-based methods. We observed that the number and quality of questions significantly impacted performance, with redundancy and conflicts in larger question sets leading to diminished results. Transfer learning approaches also provided similar, and in some cases lower performance compared to zero-shot methods that suggests the reliability and versatility of the proposed detection pipeline.

7. Acknowledgment

Computational resources for this research have been supported by the NSF National Research Platform, as part of GP-ENGINE (award OAC #2322218)

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4, 5

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. 6, 8

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3, 5, 6

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 2

[5] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 8

[6] Wen Dai, Guanghui Hu, Xin Yang, Xian-Wu Yang, Yi-Han Cheng, LiYang Xiong, Josef Strobl, and Guoan Tang. Identifying ephemeral gullies from high-resolution images and dems using flow-directional detection. *Journal of Mountain Science*, 17:3024–3038, 12 2020. 2

[7] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. 5

[8] Haonan Guo, Xin Su, Chen Wu, Bo Du, Liangpei Zhang, and Deren Li. Remote sensing chatgpt: Solving remote sensing tasks with chatgpt and visual models, 2024. 1

[9] Christos Karydas and Panos Panagos. Towards an assessment of the ephemeral gully erosion potential in greece using google earth. *Water*, 12(2), 2020. 2

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[11] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M ̃3 it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 6

[12] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[13] Boyang Liu, Ziyu Chen, Bin Li, Shufang Wu, Hao Feng, Xiaodong Gao, and Kadambot H.M. Siddique. Modeling of driving factors and headcut rates of ephemeral gullies in the loess plateau of china using high-resolution remote sensing images. *International Journal of Digital Earth*, 17(1):2369632, 2024. 2

[14] Boyang Liu, Biao Zhang, Hao Feng, Shufang Wu, Jiangtao Yang, Yufeng Zou, and Kadambot H.M. Siddique. Ephemeral gully recognition and accuracy evaluation using deep learning in the hilly and gully region of the loess plateau in china. *International Soil and Water Conservation Research*, 10(3):371–381, 2022. 1

[15] Boyang Liu, Biao Zhang, Hao Feng, Shufang Wu, Jiangtao Yang, Yufeng Zou, and Kadambot H.M. Siddique. Ephemeral gully recognition and accuracy evaluation using deep learning in the hilly and gully region of the loess plateau in china. *International Soil and Water Conservation Research*, 10(3):371–381, 2022. 2

[16] Boyang Liu, Biao Zhang, Ziming Yin, Bai Hao, Shufang Wu, Hao Feng, and Kadambot H. M. Siddique. Ephemeral gully development in the hilly and gully region of china's loess plateau. *Land Degradation & Development*, 35(1):128–141, 2024. 2

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 6

[19] Ollama. Ollama github repository. https://github.com/ollama/ollama. 6

[20] Lucas Prado Osco, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, and José Marcato Junior. The potential of visual chatgpt for remote sensing, 2023. 1

[21] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 2

[22] Jean Poesen. Soil erosion in the anthropocene: Research needs. *Earth Surface Processes and Landforms*, 43(1):64–84, 2018. 1

[23] 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, 4

[24] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[25] David A Reece, John A Lory, Timothy L Haithcoat, Brian K Gelder, and Richard Cruse. Using google earth imagery to target assessments of ephemeral gully erosion. 2023. 2

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference,*

Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015. 2

[27] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17542–17552, 2024. 2

[28] Aleksey Y Sheshukov, Lawrence Sekaluvu, and Stacy L Hutchinson. Accuracy of topographic index models at identifying ephemeral gully trajectories on agricultural fields. *Geomorphology*, 306:224–234, 2018. 2

[29] Otilia Stretcu, Edward Vendrow, Kenji Hata, Krishnamurthy Viswanathan, Vittorio Ferrari, Sasan Tavakkol, Wenlei Zhou, Aditya Avinash, Emming Luo, Neil Gordon Alldrin, et al. Agile modeling: From concept to classifier in minutes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22323–22334, 2023. 3

[30] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alphaclip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, 2024. 2

[31] Imad Eddine Toubal, Aditya Avinash, Neil Gordon Alldrin, Jan Dlabal, Wenlei Zhou, Enming Luo, Otilia Stretcu, Hao Xiong, Chun-Ta Lu, Howard Zhou, et al. Modeling collaborator: Enabling subjective vision classification with minimal human effort via llm tool-use. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17553–17563, 2024. 1, 2, 3, 4, 8

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 5, 6

[33] U.S. Department of Agriculture, Farm Service Agency. National Agriculture Imagery Program (NAIP), 2024. 4

[34] T Wolf. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 6

[35] Wenjia Xu, Zijian Yu, Yixu Wang, Jiuniu Wang, and Mugen Peng. Rs-agent: Automating remote sensing tasks through intelligent agents, 2024. 1

[36] Xin Yang, Wen Dai, Guoan Tang, and Min Li. Deriving ephemeral gullies from vhr image in loess hilly areas through directional edge detection. *ISPRS International Journal of Geo-Information*, 6(11), 2017. 2