A Multimodal Benchmark for Framing of Oil & Gas Advertising and Potential Greenwashing Detection

Gaku Morio[1*] **Harri Rowlands**[3†] **Dominik Stammbach**[4]
Christopher D. Manning[2] **Peter Henderson**[4]
[1]Hitachi, Ltd. [2]Stanford University
[3]Centre for the Acceleration of Social Technology [4]Princeton University
gaku.morio.vn@hitachi.com
manning@stanford.edu
harri@wearecast.org.uk
{dominsta,peter.henderson}@princeton.edu

Abstract

Companies spend large amounts of money on public relations campaigns to project a positive brand image. However, sometimes there is a mismatch between what they say and what they do. Oil & gas companies, for example, are accused of "greenwashing" with imagery of climate-friendly initiatives. Understanding the framing, and changes in framing, at scale can help better understand the goals and nature of public relations campaigns. To address this, we introduce a benchmark dataset of expert-annotated video ads obtained from Facebook and YouTube. The dataset provides annotations for 13 framing types for more than 50 companies or advocacy groups across 20 countries. Our dataset is especially designed for the evaluation of vision-language models (VLMs), distinguishing it from past text-only framing datasets. Baseline experiments show some promising results, while leaving room for improvement for future work: GPT-4.1 can detect environmental messages with 79% F1 score, while our best model only achieves 46% F1 score on identifying framing around green innovation. We also identify challenges that VLMs must address, such as implicit framing, handling videos of various lengths, or implicit cultural backgrounds. Our dataset contributes to research in multimodal analysis of strategic communication in the energy sector.

1 Introduction

Framing plays an important role in how people perceive and evaluate information: *"To frame is to select some aspects of a perceived reality [...], to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation [...]"* [15]. Recognizing this, oil and gas (O&G) companies have been careful in messaging their values, visions, and how they operate [see e.g., 17, 22, 42, 21]. Previous work qualitatively investigated self-portrayal and framing in O&G campaigns and climate communication [22, 39]. For example, despite their dependence on fossil fuels, companies emphasize their contributions to reducing emissions or highlight that they are essential for keeping critical infrastructure such as hospitals running [22]. These messages are considered forms of *greenwashing*,[3] as they tend to overstate corporate environmental contributions or distract from genuine climate action [13, 29].

[*]This work was undertaken in part during the author's time at Stanford University and Hitachi America, Ltd.
[†]This work was undertaken in part during the author's time at InfluenceMap CIC.
[3]Usually defined as *"behavior or activities that make people believe that a company is doing more to protect the environment than it really is"*[6].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Detection of greenwashing-related information has recently been transformed into computational benchmarks [27], e.g., a framing detection task from advertisement (ad) text [37]. These benchmarks can be used to assess and improve the capabilities of models to detect framing in public relations (PR) campaigns [37]. Accurate computational approaches have the potential to help interpret framing at scale [e.g., 35, 40, 20]. Interpreting the framing accurately has a significant potential role in social science and policymaking, especially for systematic understanding and monitoring of greenwashing.

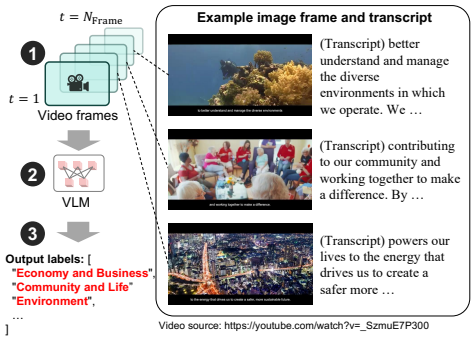

Figure 1: Overview of the task of our dataset.

However, previous related benchmarks do not consider non-text modalities, which is an important aspect of greenwashing [22, 37]. Previous qualitative analyses suggest the imagery of video ads can include strategic and misleading framings of greenwashing [22, 28]. Without understanding video-based framing, it is impossible to holistically and systematically understand corporate strategic messaging or potential patterns of greenwashing – especially if videos contain no spoken language (around 30% of the videos in our dataset, see Table 1).

Vision language models (VLMs) can be useful for detecting such video-based framing at scale [43], but there exists no evaluation or benchmark for assessing this capability. The lack of suitable benchmarking for video-based framing limits model development and potential social science applications of VLMs.

In this paper, we introduce a benchmark dataset for identifying framing techniques in O&G advertising and marketing. Figure 1 illustrates the task setting. The input is a video posted on social media by an O&G entity, such as a company or advocacy group. The output is a set of framing techniques, e.g., 'Community and Life' and 'Environment'. In our dataset, we cover major framing types for two sources: (i) The FACEBOOK domain includes video ads published by O&G companies and advocacy groups, where the videos are short and used in political campaigns to support the O&G industry. These include seven greenwashing-related framing types for 17 entities, labeled by prior work [22, 37]. (ii) The YOUTUBE domain includes marketing videos from 42 O&G related companies, where the videos are longer than those of FACEBOOK and used to convey specific impressions to viewers. We annotate six different framing types such as 'Environmental' or 'Community and Life'. By covering both domains, our dataset supports robust cross-domain evaluation of VLMs and provides a foundation for future research in energy and climate communication.

Our contributions can be summarized as follows:

(i) Multimodal dataset for framing of ads of O&G entities: We provide the first benchmark dataset of framing analysis in O&G video ads. The dataset includes fine-grained, expert-annotated labels capturing various framings across two domains. The dataset is designed to evaluate VLMs on real-world strategic framings and supports cross-domain, entity-level, and temporal analysis.
(ii) Benchmark evaluation of VLMs: We benchmark the capabilities of 6 recent VLMs on our dataset by employing zero-shot and in-context learning settings. We also test a custom-built 1-shot prompting mechanism. Overall, while our pipeline improves performance significantly, we find that VLMs still face challenges, with room for improvement to improve accuracy and consistency across other labels (for example 'Green Innovation') and cultural contexts.
(iii) Towards systematic greenwashing detection: Our benchmark can potentially be interpreted as a benchmark for detecting greenwashing. In Section 5, we discuss the relationship between greenwashing detection and our benchmark as well as a pilot study, including temporal trend analysis and company-level framing profiling for greenwashing risk assessments.

The **code** (`https://github.com/climate-nlp/multimodal-oil-gas-benchmark`) and **data** (`https://huggingface.co/datasets/climate-nlp/multimodal-oil-gas-benchmark`) are available.

2 Dataset Construction

We design the dataset considering the following needs:
(i) **Cross-domain and context-aware benchmarking:** Framings can vary across platforms. To evaluate VLMs across different contexts, we consider two domains, FACEBOOK and YOUTUBE. FACEBOOK enables us to capture framing related to climate obstruction[4] from energy companies and their value-chain entities within O&G ad campaigns. On the other hand, the YOUTUBE domain enables us to capture corporate strategies and implicit messages through official energy-related corporate channels. This dual domain design allows for the benchmarking of both generalization and domain-specific performance.
(ii) **Reliable annotations with diverse coverage:** We construct a dataset spanning 706 videos, 35k seconds of footage, 5.6k transcript segments (segments based on speech breaks extracted by Whisper-1 [36]), and 1.1k annotations – a relatively large endeavor given the challenge of gathering expert-level annotations in this area.

In addition, our dataset spans more than 50 entities across 20 countries and includes videos published from 2010 to 2025, offering diversity in cultural context and changing framing strategies over time.
(iii) **Supporting practical downstream scenarios:** The dataset is designed not only for benchmarking but also to support downstream tasks such as potential automated greenwashing detection. We release the dataset with metadata including entity information, timestamps, and URLs, enabling integration into broader research.

2.1 FACEBOOK

This subset is developed upon the previous work of Holder et al. [22] and Rowlands et al. [37]. The former [22] created a dataset to analyze Facebook ads from 2020 to 2021 in the United States, which allegedly contain messages of climate obstruction by O&G entities. The latter [37] subsequently converted the dataset into a multi-label classification task. Because their work did not mainly consider videos attached to the ads, we collected 320 videos from the original ones [22, 37] and combined these videos with the annotated labels to create a new video dataset.

Nature of Domain – Climate Obstruction Framing. Previous literature [22, 37] points out O&G companies and their agents greenwash through Facebook ads where they disseminate messages emphasizing the necessity and significance of fossil fuels, ultimately obstructing climate actions. We refer to these messages as 'climate obstruction framing' similar to existing work [37]. Specifically, Holder et al. [22] decomposed the framing into four broad categories. 'Community & Resilience' emphasizes the positive impact of the O&G industry on tax revenues and job creation. 'Green Innovation and Climate Solutions' highlights emissions reduction targets or misleadingly describes O&G as "clean". 'Pragmatism' portrays O&G as reliable and affordable, or essential raw materials for non-power-related goods such as hand sanitizer. 'Patriotic Energy Mix' stresses the importance of domestic O&G production for energy independence and security.

Rowlands et al. [37] used more fine-grained labels originally defined by Holder et al. [22], breaking down the four broader categories into seven subcategories. Because fine-grained labels are more informative, our study also adopts these labels. Below are the fine-grained label definitions (most descriptions are *quoted* from the original works [22, 37]) used in our dataset:
– **CA**: *Helps national/local economies/communities, including through philanthropic efforts.*
– **CB**: *Creates or sustains jobs.*
– **GA**: *Emissions reductions and transitioning the energy mix.*
– **GC**: *'Clean' gas as a climate solution.*
– **PA**: *Oil & gas as energy sources are a pragmatic choice and critical for maintaining functioning or optimal power systems.*
– **PB**: *Oil & gas are needed as raw materials for alternative (non-power-related) uses and manufactured goods.*
– **SA**: *The production of domestic O&G reserves benefits the US, including through energy independence or energy leadership.*

Dataset Construction. Our dataset builds on the above defined labels, while our contribution goes beyond repackaging: We are the first to align videos with the text-derived labels in this domain. We

[4] Campaigns to delay climate actions by companies and advocacy groups for example, [22].

obtained videos corresponding to each ad from the original dataset [37]. Ads without videos or ads that had already been removed were excluded. We transcribe the videos using a speech-to-text model (i.e., Whisper-1 [36]) to create the transcript text in the experiments. Each video can have up to four labels [37], making this a multi-label classification task. Previous literature [22, 37] discussed the inter-annotator agreement (IAA) of the original text-based dataset; please refer to those studies for further details. One limitation of our dataset is that the original annotations were mainly based on textual content, not videos, thus labels should be considered "distant". Despite this, our experiments show that models can still extract meaningful patterns. For further analysis of the distant labels, see Appendix A.5.2. Finally, we randomly split the dataset into training and test sets (by 50:50). We stored the dataset in a JSON Lines file (see Appendix A.5.1).

2.2 YOUTUBE

This subset consists of 386 annotated YouTube videos from major energy-related companies, capturing the "impressions" conveyed to viewers.

Nature of Domain – Framing by Impressions. YouTube has become an essential venue for companies targeting the general public with ads [16]. In our preliminary investigations, we found that energy companies strategically use YouTube to disseminate various implicit messages aimed at bolstering public perception of their industry. Initially, we attempted annotations using the same schema of FACEBOOK, but achieving a reasonable IAA was challenging due to the implicit nature of the video content. Unlike Facebook ads, YouTube videos rarely explicitly mention job creation or energy independence. They presented visual imagery, such as wind turbines, to implicitly suggest environmental commitment, or footage of workers at oil facilities to foster trust.

After rounds of annotations and discussions, we decided not to limit annotating climate obstruction framing in our dataset. Instead, we develop the following label definitions to better capture these implicit impressions, while retaining Holder et al. [22]'s high-level idea:
– **Community and Life**: Impressions of the company or O&G contributing to daily life, community, culture, transportation, sports, and charitable efforts. E.g., footage of daily life or framings about oil usage for cooking.
– **Economy and Business**: Impressions of contributing to economic prosperity, business development, or tax revenue by the company or O&G. E.g., testimony from local businesses.
– **Work**: Impressions of job creation or reliable workplaces and employees. E.g., an image of smiling workers at an oil plant.
– **Environment**: Impressions of reducing GHG emissions, supporting renewable energy, environmental responsibility, or "clean" O&G usage. E.g., an image of renewable energy like solar panels and wind turbines.
– **Green Innovation**: Impressions of innovation and visions for a green future, including the development of efficient new energy technologies. E.g., an image of a research lab, developing new climate solutions.

– **Patriotism**: Impressions of contributing positively to the country, promoting national pride, and enhancing energy independence. E.g., a framing that emphasizes national brands such as "we produce the US brand."

We conducted multiple discussions as well as input from a non-profit community to provide the above typologies. Each video can have multiple labels. Labels may be inherently overlapping (e.g., local reforestation activities can imply both 'Community and Life' and 'Environment'). Annotators were asked to follow their intuition and annotate all relevant labels.

Dataset Construction. We obtain a list of target entities by referencing the list of LobbyMap (`https://lobbymap.org/`), a platform that evaluates corporate engagement in climate policy. We initially retrieve up to 30 videos per company by searching ads on each corporate official YouTube channel, resulting in a total of 720 videos (some channels did not have 30 videos available). We wanted to collect a wide range of advertising-like content from companies, and did not want to limit videos to specific product promotions or TV commercials. At the same time, videos that were clearly not advertising-like were excluded during the annotation process, ensuring the quality of the annotated dataset. Next, we randomly sample 500 videos for annotation. Videos that are deleted or clearly not ads (e.g., earnings calls) are excluded, leaving a final set of 386 annotated videos. Annotation guidelines are refined over multiple rounds, achieving a final Fleiss' Kappa [18] agreement score

Table 1: Dataset statistics

	YouTube	Facebook
# Videos	386	320
# Transcript avail.	377	119
# Transcript segm.	4836	771
# Length (sec.)	29545	5931
# Entities	42	17
# Countries	20	1
Annotated by	Humans	Distant
Labels	802	381

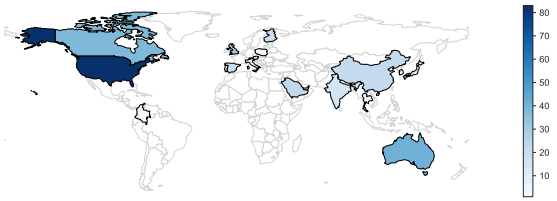

Figure 2: The country distribution (based on headquarters location) of YouTube. Note that we primarily focus on English videos produced by multinational corporations.

of 0.61.[5] Although this is not an almost perfect agreement, our labels are designed to reflect soft impressionistic framings, not hard factual typologies. This aligns with how greenwashing is typically found in real-world. Appendix A.5.4 describes details on the data collection and annotation. Finally, we transcribe the videos using Whisper-1. We randomly split the dataset into training and test sets (by 50:50). The dataset is stored in JSON Lines format (see Appendix A.5.3).

3 Dataset Analysis

Table 1 shows the statistics for videos, entity-related information, and labels. We have a total of 706 videos totaling 35,476 seconds, 1,183 annotated labels, and 5,607 transcript segments. Below we discuss the difficulty and challenges as a benchmark dataset:

– **Modality Importance:** Table 1 shows that most videos of YouTube have transcripts, where around 37% of videos do not contain transcripts in Facebook. This might be because videos on Facebook are not the primary content of the ads, and text associated with the ad typically plays a more important role. This property suggests that the importance of visual and transcript information could vary across domains. Moreover, the lack of transcripts highlights the need to incorporate the visual modality as well to improve model predictive performance. Handling both videos with transcripts and those without may present a new challenge for the model.

– **Video Length:** Table 1 shows that YouTube videos tend to be longer than those of Facebook (the length distribution is shown in Appendix Figure 9), suggesting that it could pose a challenge for VLMs, as they have to handle videos of varying lengths.

– **Entity Coverage:** Table 1 shows that YouTube and Facebook datasets contain diverse entities. All entity names are listed in Appendix A.6.2. Our dataset includes a wide range of energy-related companies and agents, ensuring that the benchmark does not overly depend on ads from a few dominant entities. This presents another challenge as models must handle differences in advertising strategies across entities. For example, companies that publicly disclose their investments in renewable energy will have different advertising strategies than those that do not.

– **Geographic Coverage:** Some entities in the Facebook dataset were multinational companies. Since this domain targets ads served in the United States, we assume that the associated country for these ads is the United States only. For YouTube, we rely on each entity's headquarters country, resulting in 20 countries in total as shown in Table 1. Figure 2 shows the distribution of countries for YouTube videos, showing that most videos are from the United States, Canada, and Australia, while also covering other major economic powers such as China and India, and oil-producing countries in the Middle East. Thus, the YouTube domain poses a challenge because VLMs must consider multiple cultural backgrounds. On the other hand, note that most of the videos we cover are English-language content and may not reflect region-specific languages or cultural aspects.

– **Imbalanced Labels:** In Appendix Table 4, we show the label distributions. The table indicates that 'Green Innovation', 'Patriotism', 'PB', 'GA', and 'SA' are low-resource labels. Interestingly, patriotic messages (i.e., 'Patriotism' and 'SA') were rare in both domains. This imbalanced property can be challenging for models, especially those relying on training data for in-context learning.

– **Temporal Coverage:** The YouTube domain covers videos from 2010 to 2025, highlighting the temporal diversity of the domain. (Please see Appendix Figure 10 for the temporal label distribution.)

[5] On the other hand, we found that this varies depending on the calculation method (e.g., see Appendix A.5.4 that also reports a 0.46 agreement score).

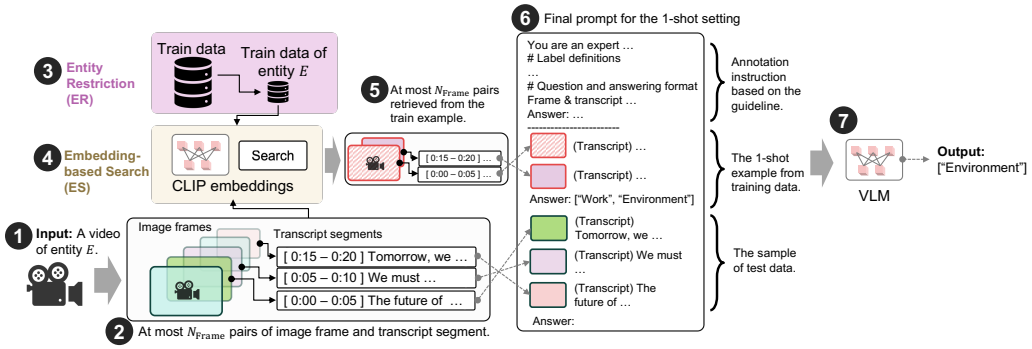

Figure 3: The overview of the entity-aware 1-shot prompt construction.

4 Benchmark Experiments

4.1 Experimental Setup

We investigate how cutting-edge VLMs perform on our two-domain datasets. Our task is a multi-label classification setting, where the input is a video and the output is a set of predicted labels (and the label set varies depending on the domain). We use the F-score to evaluate classification performance.

Since the task requires nuanced interpretation of video content, if examples are shown to VLMs, it can significantly improve classification performance. Therefore, we conduct two types of experiments: **Zero-shot setting**, where no training examples are shown, and **1-shot setting**, where one example from the training set is included in the prompt. Due to limited computational resources, we focus on the 1-shot setting, but also show K-shot results using a small model in Appendix A.7.

Models. We benchmark various cutting-edge VLMs, from open-weight models to a closed model, as follows. **DeepSeek-VL2** [48]: A relatively small (4.5B parameters) but high-performing multimodal model. **Qwen2.5-VL** [4]: A multimodal LLM based on Qwen2-VL [46] and Qwen-VL [3], designed for long video understanding, and known to outperform GPT-4o in some tasks. **InternVL2** [10]: An improved version of InternVL [11], trained with staged pretraining on large-scale vision-text datasets. **GPT-4.1** [31]: One of the latest LLMs with larger context windows. **GPT-4o-mini** [30]: A smaller variant of GPT-4o [32], which is a cost-efficient LLM that achieves near state-of-the-art performance on various multimodal tasks. For more details on the implementation, we refer to Appendix A.8.

Baseline Prompt Construction. Similar to existing literature [50, 23, 38, 26], we split each video into frame images, and then sample up to N_{Frame} to input into the VLM. At a high level, the prompt for the input consists of triplets: (i) Annotation instruction based on the guideline (c.f., Appendix Figure 12 and Figure 13), (ii) up to N_{Frame} sampled frame images, and, if available, the corresponding transcript segments, and (iii) (if 1-shot) a training example described in the same manner of (ii). The frames are dynamically selected based on the transcript. Specifically, for each transcript segment extracted by Whisper-1, we take the mean timestamp between its start and end points, and select the corresponding frame. Due to computational resource constraints, we set N_{Frame} differently depending on the model. For GPT-4.1, GPT-4o-mini, and Qwen2.5-VL, we set $N_{\text{Frame}} = 10$. For InternVL2 and DeepSeek-VL2, we set $N_{\text{Frame}} = 3$. We also instruct the VLM to output the answer as a JSON list of labels. An overview of the prompt construction and frame-transcript pairing can be found in part of Figure 3 (see (1), (2), (6), and (7) of the figure).

Entity-aware 1-shot Prompt Construction. As mentioned in Section 3, video characteristics can differ significantly across entities. This will necessitate selecting an informative sample for 1-shot prompting as much as possible. To verify this, we provide an entity-aware prompt construction approach for the 1-shot setting. The high-level idea of the approach is to select the 1-shot sample based on entity-aware similarity search from the training data, as is illustrated in Figure 3. The similarity search was inspired by a study that utilizes CLIP embeddings for retrieval-based prompting [24] in an image and text classification task. We adapt the work into our video classification task by using both frame images and transcript segments for the embeddings.

Table 2: Zero-shot and 1-shot experimental results in F-scores (%). 'All' denotes the micro-averaged score for all labels.

Model		YouTube							Facebook							
		All	Comm. &Life	Work Env.	Green Innov.	Econ. &Bus.	Patrio -tism		All	CA	CB	PA	PB	GA	GC	SA
Zero-shot																
DeepSeekVL2	4.5B	45.3	67.2	29.3 50.0	0.0	36.5	0.0		23.2	26.8	23.0	23.6	33.3	16.6	5.8	28.5
InternVL2	8B	53.1	71.7	55.2 50.0	35.2	39.7	27.2		22.5	31.2	6.4	8.6	4.8	37.6	11.4	31.8
Qwen2.5-VL	7B	37.3	32.0	42.1 37.6	29.0	48.7	28.5		25.4	32.5	40.0	19.1	14.8	29.6	20.0	23.5
Qwen2.5-VL	32B	60.7	68.8	61.7 73.5	**46.9**	47.3	42.8		49.0	42.8	78.6	35.0	**60.0**	46.6	38.0	**51.4**
GPT-4o-mini	-	60.5	72.4	66.2 72.8	39.2	41.1	43.1		54.2	40.9	**79.3**	56.5	40.0	42.5	52.1	43.9
GPT-4.1	-	**71.0**	**84.9**	**77.3 79.4**	46.1	**52.9**	**52.1**		**61.1**	**48.2**	73.5	**67.8**	42.8	**50.0**	**76.3**	39.0
1-shot																
DeepSeekVL2	4.5B	49.7	68.6	49.3 47.3	21.4	36.0	32.2		62.3	47.3	51.7	65.5	**61.5**	63.1	78.8	51.2
InternVL2	8B	56.7	78.1	61.9 53.0	34.1	39.3	25.0		46.2	30.0	29.2	60.5	35.2	41.2	71.6	35.0
Qwen2.5-VL	7B	59.2	67.3	62.4 65.6	**46.5**	47.7	43.9		58.2	38.4	62.5	64.7	33.3	**66.6**	75.0	44.4
Qwen2.5-VL	32B	66.2	76.0	70.2 77.4	45.8	48.9	48.2		70.5	44.8	**85.2**	67.2	54.5	**66.6**	**90.3**	**64.5**
GPT-4o-mini	-	63.0	72.9	68.6 74.9	39.9	45.2	51.2		65.2	53.3	80.6	67.2	**61.5**	52.3	74.0	51.2
GPT-4.1	-	**69.3**	**80.6**	**75.0 78.3**	41.6	**52.1**	**59.0**		**72.6**	**60.3**	81.2	**76.1**	54.5	63.4	81.9	62.5

Concretely, given a test video from entity E, our method proceeds as follows: (i) **Entity Restriction (ER)** restricts the pool of candidate training videos to those belonging to the same entity E. This ensures that the examples used for in-context learning are relevant to the characteristics of the target entity. (ii) **Embedding-based Search (ES)** performs a similarity search [24] over the restricted training set using embeddings. First, the frame images and transcript segments are embedded by CLIP: $\mathbf{e}_{\text{Frame}} = \frac{1}{N_{\text{Frame}}} \sum_{i=1}^{N_{\text{Frame}}} \mathbf{e}_{\text{Frame-}i}$ and $\mathbf{e}_{\text{Transcript}} = \frac{1}{N_{\text{Frame}}} \sum_{i=1}^{N_{\text{Frame}}} \mathbf{e}_{\text{Transcript-}i}$. Here, $\mathbf{e}_{\text{Frame-}i}$ and $\mathbf{e}_{\text{Transcript-}i}$ represents the i-th frame image embedding vector and transcript segment embedding vector, respectively. The video representation is then computed as $\mathbf{e} = \lambda_{\text{Frame}}\mathbf{e}_{\text{Frame}} + \lambda_{\text{Transcript}}\mathbf{e}_{\text{Transcript}}$. We set the hyperparameters $\lambda_{\text{Frame}} = 0.5$ and $\lambda_{\text{Transcript}} = 0.5$ through experiments. Finally, the most similar training video is retrieved based on cosine similarity. The rest of the process for constructing the prompt is the same as the 1-shot process for the baseline prompt. This approach makes VLMs see an example that is not only similar in content but also aligned with the features of the target entity.

4.2 Results and Discussion

Overall Results. Table 2 reports the experimental results on YouTube and Facebook domains. Each table shows F-scores along with label-wise scores for both zero-shot and 1-shot settings. 'All' indicates the micro-averaged F-score across all labels.

In YouTube, GPT-4.1 performed best while GPT-4o-mini and Qwen2.5-VL (32B) achieved reasonably high F-scores in both zero-shot and 1-shot settings. Given the nuanced nature of our annotation task and the IAA of 0.61, these results are promising. We observed a large performance gap between 1-shot and zero-shot settings, especially for open-weight models such as Qwen2.5-VL. This suggests that providing a suitable example in the prompt is important. The exception is GPT-4.1 where the zero-shot setting outperformed the 1-shot setting. This demonstrates the remarkable capability of GPT-4.1 in this domain. For the label-wise performance, labels such as 'Community and Life', 'Work', and 'Environment' were classified with relatively higher accuracy. These labels often correlate with clear visual or textual cues, such as images of families, workers, or mentions of environmental commitments. In contrast, low-resource labels like 'Patriotism', 'Economy and Business', and 'Green Innovation' may be difficult for models to classify. This could be because the messages rely more on subjective and subtle interpretation.

In Facebook, a similar trend was observed, while the 1-shot results consistently outperformed the zero-shot results across all models with a larger performance gap. VLMs may struggle to classify the fine-grained labels, suggesting the importance of showing the 1-shot example in the prompt. Interestingly, despite its smaller size, DeepSeekVL2 performed similarly to GPT-4o-mini in the 1-shot setting. This suggests that even smaller-scale VLMs can be effective under certain conditions. The

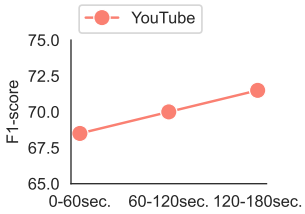

Figure 4: The video length and F-score of GPT-4.1.

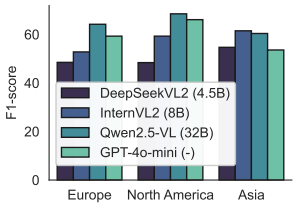

Figure 5: The region and F-score for YouTube. We exclude Middle East from Asia.

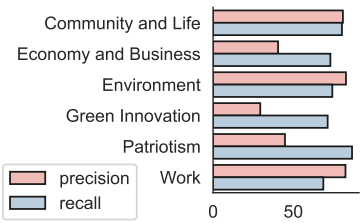

Figure 6: The precision and recall analysis for GPT-4.1 in YouTube.

label-wise F-scores show that labels such as 'PB' showed lower F-scores across models, suggesting that these categories are particularly difficult for the models to interpret.

Below, we further discuss ablation studies and error analyses. Otherwise specified, we use the 1-shot prompting with ES and ER.

Ablation – Modality Importance. Table 3 shows the results of ablation study using Qwen2.5-VL 32B. The '**T** $=\times$' (i.e., the 2nd row of the table) means that the transcript input is ablated. In YouTube, we obtained 61.2% F-score, while in Facebook we obtained 60.6%, suggesting the modality from the transcript of the video is important for the inference of VLMs. There is a huge gap between models with and without transcript input, especially in Facebook. On the other hand, smaller models (e.g., DeepSeekVL2) on YouTube, including transcript inputs sometimes degrades performance (see Table 8). These results above align with our idea in Section 3 that the importance of visual and transcript information varies across domains.

Table 3: Ablation results in the 1-shot prediction. **T** represents the transcript input, **ES** represents the embedding search in the 1-shot sample selection, and **ER** represents the entity restriction in the 1-shot sample selection. The full results for all models can be found in Table 8.

Model	T	ES	ER	YouTube	Facebook
Qwen2.5-VL 32B	✓	✓	✓	**66.2**	**70.5**
	×	✓	✓	61.2	60.6
	✓	×	✓	64.0	59.1
	✓	✓	×	65.6	68.1

Ablation – Entity Restriction and Embedding-based Search. Table 3 also shows the effect of the entity-aware 1-shot prompt construction (see **ES** and **ER** in the table). Incorporating both components (ES, ER) generally led to better performance of Qwen2.5-VL 32B predictions, highlighting the importance of retrieving similar examples for the 1-shot prompt. However, the effect of ER was mixed when observing all models (see Appendix Table 8). When only a few training videos per entity are available, restricting the search space may limit the quality of the retrieved examples.

Error Analysis – Video Length. Figure 4 shows the F-scores for different video lengths. Interestingly, shorter videos (i.e., 0–60 sec.) seem to be more challenging for the model. This might be because shorter videos are more contextualized and vague, making it difficult for models to predict framings. The result suggests shorter videos need careful handling.

Error Analysis – Geographic Effect on Models. Figure 5 shows the F-scores in each geographical region (obtained from the countries where the headquarters are located) for YouTube. The figure shows GPT-4o-mini and Qwen2.5-VL perform well on videos from European and North American companies, while DeepSeekVL2 and InternVL2 outperform GPT-4o-mini for videos from Asia. The result suggests that each VLM has a different ability to handle specific cultural contexts. This insight supports our idea in Section 3 that VLMs must consider multiple cultural backgrounds.

Error Analysis – Over-label or Under-label. Figure 6 shows a precision and recall analysis across different labels for YouTube to investigate the extent to which a VLM over- or under-labels. For labels with relatively higher overall F-scores, such as 'Community and Life', 'Environment', and 'Work', precision tends to be higher than recall. For more difficult labels such as 'Economy and Business', 'Green Innovation', and 'Patriotism', recall tends to be higher than precision. This suggests that for difficult labels, VLMs tend to over-label, while experts tend to under-label when assigning such labels.

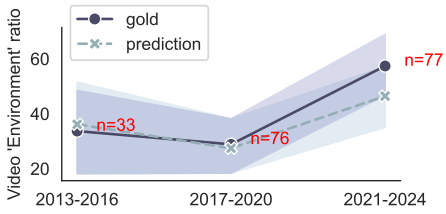

Figure 7: The ratio of videos with 'Environment' on YouTube by GPT-4.1. We show the 95% confidence interval with bootstrap resampling.

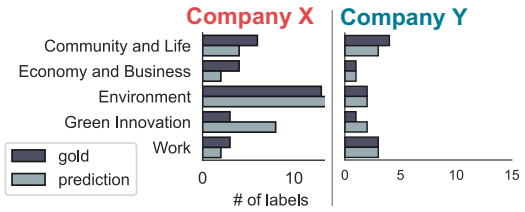

Figure 8: The GPT-4.1 predicted and gold labels for two example companies.

We also note significant variation in the labelling habits of different models. We investigated the number of output labels by the zero-shot prompting. We found Qwen2.5-VL 7B is by far the most under-labeler (171 labeled, mean: 349) while GPT-4o-mini is an over-labeling model (409 labeled). On the other hand, models struggled to capture the most common co-occurrences demonstrated in the gold dataset. DeepSeekVL2 shows a strong pairing of the labels 'Community and Life' and 'Economy and Business' (115 times vs. 63 times in the gold), despite this being rare in the gold. Meanwhile, DeepSeekVL2 does not accurately reflect the most common pair in the gold dataset – 'Community and Life' and 'Work' (DeepSeekVL2: 17 vs gold: 131).

Takeaway. The above discussions identify challenges that VLMs must address, such as handling different video lengths, over-label or under-label annotations, and cultural backgrounds. Through benchmark evaluations using our dataset, researchers can select the VLM that best suits their needs, recognize the limitations of each model, and test new methods.

5 Discussion: Towards Multimodal Greenwashing Detection

Videos in our benchmark are potentially candidates for greenwashing. The framing of Facebook videos in our dataset represents the following forms of greenwashing: Influencing the general public to perceive O&G as "clean" (i.e., 'GC') or emphasizing the necessity of oil enhances welfare for communities or is a pragmatic necessity (i.e., 'CA', 'CB', 'PA' and 'PB'). This emphasizes the necessity of the O&G industry and reduces awareness of environmental impact, which is related to obstruction and greenwashing by selective disclosure [13, 29]. Our benchmark allows evaluating whether VLMs can accurately detect such forms of greenwashing.

The framing of 'Environment' and 'Green Innovation' in YouTube often leverages vagueness to portray an environmentally friendly image without making specific commitments [29, 41]. Analogously, the labels 'Community and Life', 'Economy and Business', and 'Work' relate to obstruction and selective disclosure similarly to the Facebook labels 'CA', 'CB', 'PA' and 'PB'.

Furthermore, we introduce the following pilot studies that assess greenwashing risks through a comprehensive analysis:

Temporal Trend Analysis. Understanding how environmental framing has shifted over time can provide valuable insights in O&G communication strategies. To this extent, we analyze the temporal trend of environment-related messaging in Figure 7. We calculate the environment label ratio, defined as the number of videos labeled as 'Environment' divided by the total number of videos, for each year. We observe that the predicted trend closely follows the ground-truth trend, suggesting that our model can reasonably capture temporal trends in framing. Importantly, the model captures the increasing trend after 2020. This correlates to the period of the Biden administration (although we do not verify its causality). Given that greenwashing typically involves emphasizing positive environmental communication [13], this increase could be interpreted as a potential signal of industry-wide greenwashing.

Company-level Analysis. Here, we present two case studies based on selected companies. Figure 8 shows the gold and predicted label distributions for Company X and Company Y (anonymized), suggesting that the model can reasonably replicate the gold distribution. Interestingly, Company X exhibits a high proportion of 'Environment' labels, showing that the company may actively use its

YouTube channel to promote a strong environmental image that exceeds the industry average. In contrast, Company Y places more emphasis on 'Community and Life', reflecting a perspective that presents O&G as essential to our daily life. Also, we found this company frequently includes other labels such as 'Work' and 'Environment' in the same videos. We observed that in such multi-labeled videos, environment-related content was often vague and embedded as part of the overall impression, rather than linked to concrete environmental performance.

In our view, the identified cases above are promising candidates to further evaluate manually. Gathering such insights through computational assistance can reduce the amount of manual research required to reach such conclusions. Further, they show the practical potential of our benchmark dataset in greenwashing detection, corporate strategy profiling, and social science research.

6 Conclusion

We propose a video classification task for framing in ad videos by O&G entities, and we introduce a benchmark dataset that consists of two sources, namely YOUTUBE and FACEBOOK. The dataset is challenging for VLMs and we see lots of room for improvement going forward.

Furthermore, our work enables practical applications such as computer-assisted multimodal greenwashing analysis or social science work about temporal trends in O&G messaging.

We hope this dataset encourages researchers and practitioners to study corporate messaging from both visual and textual perspectives. Future directions include extending the task to more granular labels, e.g., whether the messaging is implicit or explicit or whether the message is portrayed visually, via spoken language and/or as captions. Lastly, we plan to include more domains with high greenwashing risks.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous referees for their helpful comments on this paper. GM conducted this research as part of the Stanford Data Applications Affiliates Program. At the time the research was conducted, GM was an employee of Hitachi America, Ltd., which provided financial support for the study.

References

[1] Anthropic. 2024. Claude 3 model family: Opus, Sonnet, Haiku. Released March 4, 2024; includes Claude 3 Opus, Sonnet, and Haiku models with advanced multimodal capabilities and large context windows.

[2] Arnav Arora, Srishti Yadav, Maria Antoniak, Serge Belongie, and Isabelle Augenstein. 2025. Multi-modal framing analysis of news. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31531–31553, Suzhou, China. Association for Computational Linguistics.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

[5] Benjamin D Blair and Larkin McCormack. 2016. Applying the narrative policy framework to the issues surrounding hydraulic fracturing within the news media: A research note. *Research & Politics*, 3(1):2053168016628334.

[6] Cambridge Dictionary. 2023. Meaning of greenwashing in English. https://dictionary.cambridge.org/us/dictionary/english/greenwashing. Accessed: 2023-05-11.

[7] Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

[8] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pages 53168–53197. Curran Associates, Inc.

[9] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems*, volume 37, pages 28662–28673. Curran Associates, Inc.

[10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Intern VL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

[13] Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32:1–12.

[14] Nick J Enfield. 2024. *Language vs. reality: Why language is good for lawyers and bad for scientists*. MIT Press.

[15] Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

[16] Mohamad Trio Febriyantoro. 2020. Exploring youtube marketing communication: Brand awareness, brand image and purchase intention in the millennial generation. *Cogent Business & Management*, 7(1):1787733.

[17] George Ferns, Kenneth Amaeshi, and Aliette Lambert. 2019. Drilling their own graves: How the European oil and gas supermajors avoid sustainability tensions through mythmaking. *Journal of Business Ethics*, 158:201–231.

[18] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

[20] Kai Gehring and Matteo Grigoletto. 2023. Analyzing climate change policy narratives with the character-role narrative framework. Working Paper 10429, CESifo Working Paper.

[21] Giuliana Gentile and Joyeeta Gupta. 2025. Orchestrating the narrative: The role of fossil fuel companies in delaying the energy transition. *Renewable and Sustainable Energy Reviews*, 212:115359.

[22] Faye Holder, Sanober Mirza, Namson-Ngo-Lee, Jake Carbone, and Ruth E. McKie. 2023. Climate obstruction and facebook advertising: how a sample of climate obstruction organizations use social media to disseminate discourses of delay. *Climatic Change*, 176(2):16.

[23] Zi-Yuan Hu, Yiwu Zhong, Shijia Huang, Michael Lyu, and Liwei Wang. 2024. Enhancing temporal modeling of video LLMs via time gating. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2845–2856, Miami, Florida, USA. Association for Computational Linguistics.

[24] Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with LMM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293, Miami, Florida, USA. Association for Computational Linguistics.

[25] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[26] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

[27] Gaku Morio and Christopher D Manning. 2023. An NLP benchmark dataset for assessing corporate climate policy engagement. In *Advances in Neural Information Processing Systems*, volume 36, pages 39678–39702. Curran Associates, Inc.

[28] Xavier Muller. 2025. People or profit? a content analysis of energy corporation's use of greenwashing in advertisements. *The Sociological Imagination: Undergraduate Journal*, 10(1).

[29] Noémi Nemes, Stephen J. Scanlan, Pete Smith, Tone Smith, Melissa Aronczyk, Stephanie Hill, Simon L. Lewis, A. Wren Montgomery, Francesco N. Tubiello, and Doreen Stabinsky. 2022. An integrated framework to assess greenwashing. *Sustainability*, 14(8).

[30] OpenAI. 2024. GPT-4o Mini. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Released July 18, 2024.

[31] OpenAI. 2025. GPT-4.1. https://openai.com/index/gpt-4-1/. Released April 14, 2025.

[32] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[34] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, joseph heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, and 5 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pages 42748–42761. Curran Associates, Inc.

[35] Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

[37] Harri Rowlands, Gaku Morio, Dylan Tanner, and Christopher Manning. 2024. Predicting narratives of climate obstruction in social media advertising. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5547–5558, Bangkok, Thailand. Association for Computational Linguistics.

[38] Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. 2024. TraveLER: A modular multi-LMM agent framework for video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9740–9766, Miami, Florida, USA. Association for Computational Linguistics.

[39] Yutong Si, Dipa Desai, Diana Bozhilova, Sheila Puffer, and Jennie C Stephens. 2023. Fossil fuel companies' climate communication strategies: Industry messaging on renewables and natural gas. *Energy Research & Social Science*, 98:103028.

[40] Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

[41] Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

[42] Geoffrey Supran and Naomi Oreskes. 2021. Rhetoric and frame analysis of ExxonMobil's climate change communications. *One Earth*, 4(5):696–719.

[43] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.

[44] Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

[45] Jiawen Wang, Longfei Zuo, Siyao Peng, and Barbara Plank. 2024. MultiClimate: Multimodal stance detection on climate change videos. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 315–326, Miami, Florida, USA. Association for Computational Linguistics.

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

[47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

[48] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

[49] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

[50] Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. 2024. OmAgent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10031–10045, Miami, Florida, USA. Association for Computational Linguistics.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions are the creation of datasets, analysis, benchmark experiments, and discussions, which are appropriately documented in the main text.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A.2.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not contain theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Major implementation details and hyperparameters are described in Section 4.1 and Appendix A.8.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We made our dataset and code public.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.1 and Appendix A.8.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In Figure 7, we report statistics suitably.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A.9.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Appendix A.1 and Appendix A.12.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix A.1.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Risk of misuse is low.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix A.12, Appendix A.8, and Section 4.1.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Appendix A.12.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components. We used an LLM for writing and editing purposes. Though this is not required to disclose, we did in Appendix A.11.

A Appendix

A.1 Ethics Statement

In alignment with the Code of Ethics guidelines [6], the following discuss ethical considerations and potential societal impact of this work.

Privacy. Videos in our dataset are voluntarily made publicly available by its producers as marketing and promotional videos. Thus, we do not see any violation of privacy issues.

Consent. The videos used in our study are publicly available and of public interest. Given the public interest about framing strategies for O&G video, it is fair use to annotate and release such videos for research purposes and do not require consent.

Deprecated Datasets. Not applicable.

Copyright and Fair Use. Given the public interest about framing strategies for O&G video, it is fair use to annotate and release such videos for research purposes.

Representative Evaluation Practice. We discuss the geographic diversity and temporal coverage of our dataset in our paper, and showcase the companies included (Table 5 and Table 7) and their headquarters' location in Figure 2. Our dataset, although limited in size, exhibits extensive coverage of large O&G companies in various countries across four continents and 15 years. Nevertheless, there exists a bias towards promotional videos of the largest O&G companies and English videos. In future work, we are excited to explore videos from smaller companies, short video formats (e.g., Facebook Reels, TikTok, YouTube Shorts) and non-English videos.

Safety. We do not see any major safety concerns in releasing annotations for promotional videos of O&G videos for educational and research purposes.

Security. We do not see any major security concerns in releasing annotations for promotional videos of O&G videos for educational and research purposes.

Discrimination. We do not see any major discrimination concerns in releasing annotations for promotional videos of O&G videos for educational and research purposes.

Surveillance. We do not see any major surveillance concerns in releasing annotations for promotional videos of O&G videos for educational and research purposes.

Deception & Harassment. We do not see any major deception & harrasment concerns in releasing annotations for promotional videos of O&G videos for educational and research purposes.

Environment. We recognize the environmental impact, e.g., energy and/or water consumption while benchmarking various models on the released dataset.

Human Rights. Not applicable.

Bias and fairness. Due to annotation artifacts and the nature of distributions of large O&G companies, our dataset contains biases related to countries and regions. Researchers employing the dataset should be aware of such biases.

A.2 Limitations

The data acquisition process includes implicit or unrecognized biases. For example, the subset from FACEBOOK was obtained dataset of the previous literature. The data collection process for the subset of YOUTUBE includes the channel search to access to relevant videos, which may not reproduce depending on the internal recommendation algorithm of the platform. We do not collect all videos but sampled from the available video set. We do not consider videos which were removed or unable to download.

Our dataset was sourced from Facebook and YouTube, a decision driven by the accessibility for data acquisition. Thus, our findings may not generalize to the broader video-based ads across diverse platforms. This study does not address the domain of short-form video content (e.g., Facebook Reels, TikTok, YouTube Shorts), despite its significant and increasing prevalence. Furthermore, our dataset

[6] https://neurips.cc/public/EthicsGuidelines

can contain a bias towards content from large, multinational corporations and advertisers in economically dominant regions or oil producing countries. Future research would benefit from incorporating more diverse domains that include emerging markets, smaller firms, and underrepresented video formats.

Our dataset has been created with best efforts in terms of scale and annotation accuracy. However, model bias may arise due to inaccurate annotations or small sample sizes. Additionally, due to the nature of advertising videos, highly similar content may be included in both the training and test data, which could potentially lead to optimistic benchmark results. Transcripts are automatically generated and may contain errors in speech recognition. Furthermore, the dataset was created primarily with videos that are understandable to English or Japanese speakers, so linguistic representativeness is limited. FACEBOOK videos are originally labeled mainly based on ad text rather than the video's overall content. This "distant annotation" approach could ignore visual and nuanced cues within the video, potentially introducing bias into model training and evaluation.

Despite spending days debating interpretations and resolving disagreements for the YOUTUBE domain, we can encounter subjective interpretations when labeling. Certain framing categories (e.g., 'Green innovation' and 'Environment') may be interpreted as overlapping in some cases. This kind of potential ambiguity may introduce additional uncertainty into model development and evaluation.

Given the size of our dataset, the generalizability of our analyses and experimental results can be restricted. Given above, our benchmark is unlikely to be suitable for fine-tuning models on the task. However, we note that there is limited availability of datasets for multi-modal analysis more broadly across all climate change domains. Even though it is not directly related to our work, MultiClimate [45], for example, is a recent dataset on climate stance detection from videos, containing a total of 100 videos. More importantly, our work is aimed at covering a diverse set of videos rather than a large number of videos. Our work is characterized by fine-grained labels across two different domains. Moreover, within each domain, we consider a variety of entities.

From a technical standpoint, our study is subject to several limitations. Computational resource constraints restricted the scale of our frame-level analysis (c.f., the number of sampled image frames input into the VLMs), potentially limiting our ability to discern highly granular temporal patterns within video content. Secondly, we only tried the single run for each VLMs. There can be statistical fluctuations for the results.

A.3 Code and Data Availability

The code and data are available on Github nd Huggingface.

A.4 Related Work

Recent progress in capabilities to analyze videos in open and closed source models make this work possible. Among others, these include GPT-4o-mini [30], DeepSeek-VL2, [48], InternVL2 [10] and the Qwen2.5-VL model family [4] and we benchmark all of these in our work. Many of the more recently released next-generation language models are capable of processing videos, e.g., Claude and Gemini and Qwen3 [1, 12, 49] and we expect this trend to continue and models to become better at analyzing videos in the near future. We leave it to future work to examine such models on our benchmark more carefully.

Further, our work connects to narratives and framing in social sciences and economics [5, 22, 14] and to work on computational narrative analysis [35, 40], framing [7, 2] and agenda setting [44]. Specifically in the environmental context, we find work facilitating the automatic analysis of company reports to extract quantities of interest, such as environmental claims [41], environmental narratives [20, 37] and corporate climate policy engagement [27]. All these studies hint at the feasibility of computational assistance for undertaking large scale data analysis for environmental narratives and framing. Our work is largely inspired by Holder et al. [22] and Rowland et al., [37] who released work similar to ours, but focused on natural language processing and textual data in O&G marketing ads.

Lastly, related work on video datasets mostly emphasizes benchmarking and assessing model capabilities [e.g., 34, 8, 9]. While we highly appreciate such work, this is only partially the focus of our paper. More importantly, we view our dataset artifact as a tool to evaluate and improve video

processing models. This in turn can facilitate computational narrative and framing analysis, and can provide computational assistance for large-scale social science studies, and finally to measure O&G framing over time.

A.5 Dataset Details

A.5.1 Detail of FACEBOOK

We stored the dataset in a JSON Lines file, where each line corresponds to a video, including not only labels but also metadata such as ID, URL, entity name, and video length in seconds. Video resolution varies (e.g., 398x224, 400x400) based on the original content, and all videos are in MP4 format, including audio (though some videos lack original audio tracks). The following is an example object of each line of the file:


```
{
  "video_id": "video_001",
  "video_url": "[ANONYMIZED]",
  "labels": ["PA"],
  "video_length_seconds": 15,
  "entity_name": "[ANONYMIZED]"
}
```


Note that anonymization is applied only in this paper; the actual released dataset maintains the original information.

A.5.2 Evaluation of Distant Labels of the FACEBOOK Domain

In the FACEBOOK domain, there may be cases where video content and labels do not perfectly align, because the original labels were created mainly for textual content. To evaluate this, one of the authors manually annotated 20 randomly selected videos based only on videos, comparing them with the original labels. F-score between the original annotations and our manual video-based annotations was 83%, suggesting that the distant labels are still reasonable quality.

A.5.3 Detail of YOUTUBE

We stored the dataset in a JSON Lines file, with each line containing metadata including video ID, URL, published date, entity name, video length in seconds, entity's headquarter country and region, channel name and ID, and video view count, in addition to annotated labels. The following is an example object of each line of the file:


```
{
  "video_id": "video_101",
  "video_url": "[ANONYMIZED]",
  "labels": ["Economy and Business", "Work",
             "Environment", "Green Innovation"],
  "video_publish_date": "2021-11-11",
  "video_title": "[ANONYMIZED]",
  "video_length_seconds": 79,
  "video_views": 12000,
  "entity_name": "[ANONYMIZED]",
  "entity_country": "[ANONYMIZED]",
  "entity_region": "[ANONYMIZED]",
  "channel_id": "[ANONYMIZED]",
  "channel_name": "[ANONYMIZED]"
}
```


Note that anonymization is applied only in this paper; the actual released dataset maintains the original information.

A.5.4 Construction Detail of YOUTUBE

As described before, we initially retrieved up to 30 videos per company by searching ads. The number of 30 results from the restriction of the search results by a Python library we used. We use 'advertisment' as a searching query, where it contains unintentional spelling mistakes. We, however,

Table 4: The label distribution

	YouTube	Facebook
Label (Train/Test)	Comm.&Life (125/122) Work (102/96) Env. (84/78) Econ.&Bus. (53/41) Green Innov. (39/28) Patriotism (19/15)	CA (27/23) CB (33/29) PA (55/64) PB (10/7) GA (15/21) GC (30/33) SA (21/13)

found many of the results by the query and the correct query 'advertisement" are identical. Although channel-search introduces potential biases, our priority was collecting advertisement-like videos.

To apply Fleiss' kappa to the multi-label task, the inter-annotator agreement was calculated by counting whether a label exists for each label, and the final score was computed. In the initial round of the guideline refining process, three annotators achieved an agreement score of 0.35 on 16 randomly selected videos. After guideline revisions, two annotators reached a 0.59 agreement score on a new set of randomly selected videos in the second round. Following further guideline improvements, the final round achieved a 0.61 agreement score between two annotators on 21 additional videos. On the other hand, calculating the kappa for each label and taking the average yielded 0.46. The 'Patriotism' label was excluded from the calculation because it lacked a label. Additionally, the 'Work' label had a kappa of -0.05, which appears to stem from an annotator's careless mistake. Although the calculation of inter-annotator agreement is subject to such constraints, GPT-4.1's F-score exceeds 70% on average, suggesting that relatively reliable annotations have been performed.

These two annotators then completed the annotation for all 386 videos using this finalized guideline. Note that, we initially attempted to simultaneously annotate whether framing was explicit or implicit. However, this was not adopted in the final annotation due to a lack of agreement levels.

A.6 Annotation Examples

Here we show some annotation examples in our guideline for the YouTube domain. Note that the following annotation examples are based on the annotator's impressions and do not represent any particular position.

Video URL: https://www.youtube.com/watch?v=xRXA9xR_8o8**.** This video gives the viewer the impression that energy is necessary to support industry and everyday life. It also includes an environmentally friendly image with a narrative "minimal environmental impact". It also includes images of workers. Finally, the labels can be 'Economy and Business', 'Work', 'Environment', and 'Community and Life'.

Video URL: https://www.youtube.com/watch?v=mVRC08LXfg0**.** This video gives the viewer the impression that the company supports charitable efforts (Community and Life). As an implicit message, the company is supporting climate actions. Finally, the labels can be 'Community and Life' and 'Environment'.

A.6.1 Further Analysis

The label distribution can be found in Table 4.

The video length distribution can be found in Figure 9.

The temporal label distribution can be found in Figure 10.

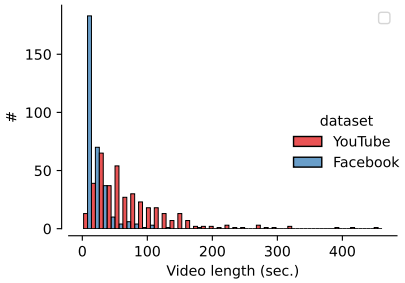

Figure 9: The video length distribution.

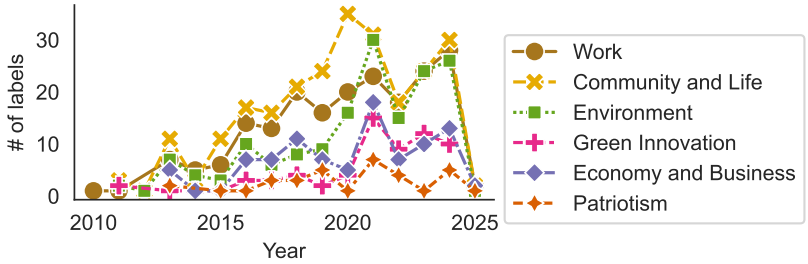

Figure 10: The temporal distribution of YOUTUBE.

A.6.2 Full Entity List

Table 5 and Table 7 shows the entity list of the dataset from FACEBOOK and YOUTUBE, respectively.

A.7 Effect of Few-shot Size

We investigate the effect of increasing the few-shot size (K) using Qwen2.5-VL 7B. We set $K = 1, 4, 8$ and evaluate the F-scores for both YOUTUBE and FACEBOOK. To secure enough search space for selecting few-shot samples, we did not use ER in this experiment. Figure 11 shows the general trend that increasing K improves F-scores. Increasing the few-shot size takes longer inference time and higher computational cost, which is the limitation resulted from the trade-off between performance and efficiency.

A.8 Implementation Detail

We use PyTorch 2.6.0 [33], HuggingFace Transformers 4.51.2 [47], and vLLM 0.8.3 [25] libraries for the model and inference implementation. For GPT-4.1, we use the version of 'gpt-4.1-2025-04-14'. For GPT-4o-mini, we use the version of 'gpt-4o-mini-2024-07-18'. For Qwen2.5-VL, we use the versions of 'Qwen/Qwen2.5-VL-7B-Instruct' and 'Qwen/Qwen2.5-VL-32B-Instruct'. For InternVL2, we use the version of 'OpenGVLab/InternVL2-8B'. For DeepSeekVL2, we use the version of 'deepseek-ai/deepseek-vl2'. For the CLIP embedding, we use 'openai/clip-vit-base-patch32' available on Huggingface. This study does not conduct any hyperparameter search and model selection. For all VLMs, we set temperature at 0. Due to limited computational resources of us, we only conduct the single run and report the F-score for each VLM.

Figure 12 and Figure 13 shows the beginning of the prompt (i.e., the annotation guideline part) of FACEBOOK and YOUTUBE, respectively.

Detail of Image Frames Extraction Transcript segments with very short text (three words or fewer) are excluded, as we found they tend to be noisy. If the number of segments exceeds N_{Frame}, excess segments and their corresponding frames are discarded. If no transcript is available (i.e., $N_{\text{Frame}} = 0$), we sample N_{Frame} frames uniformly from the video. In this case, the corresponding transcript segments are set as "N/A".

Table 5: Entity list of FACEBOOK

Name	# Video
AGA	35
API	50
Alliancefor_MI	4
BP	7
CA_forAffordableandReliableEnergy	1
CA_forEnergyIndependence	29
CO_forResponsibleEnergyDevelopment	2
ConsumerEnergy	4
Enbridge	4
EnergyTransfer	40
ExxonMobil	33
GreatLakes_MI	4
NMOGA	48
OH_Oil&Gas	5
Partnership_EnergyProgress	2
TX_Oil&Gas	3
Williams	49

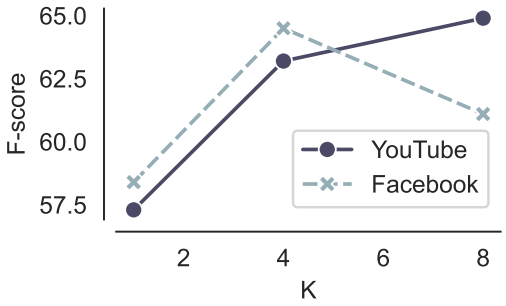

Figure 11: The effect of few-shot size for Qwen2.5-VL 7B.

A.9 Computational Resources

We use Amazon EC2's 'g5.48xlarge' (8 x A10G Tensor Core GPU (192 GiB GPU memory) and 768 GiB memory) for experiments using open-source VLMs. The experiment was conducted over a period of about two days.

A.10 Full Results of the Ablation Study

Table 8 shows the full results of the ablation study.

A.11 Declaration of LLM Usage

We used ChatGPT and Gemini 2.5 in parts of our paper to translate, correct grammar, and polish the writing.

```
As an expert annotator, please label the following video advert according to the described typology. Many adverts will not be relevant so please label them as X. We are looking for narratives specifically from the oil and gas sector.

Community & Resilience
  CA: Emphasizes how the oil and gas sector contributes to local and national economies through tax revenues, charitable efforts, and support for local businesses.
  # Here's the annotation criteria points:
    (General)
      - the O&G sector or an O&G company contributing to national or local GDP.
      - helping local communities or society through tax revenues from O&G (often stated that these are used to build schools, hospitals, roads etc.)
      - helping local businesses (typically through indirect spending).
    (Corresponding negative statements including that without the O&G sector/company)
      - the economy would suffer.
      - local communities would suffer due to the loss of tax revenue.
      - local businesses would suffer due to the lack of indirect spending.
    (Philanthropic efforts)
      - the O&G sector/company is participating in voluntary efforts (non-climate related).
      - donating money to charitable causes.
      - donating money to fund or train first responders.
      - donating supplies.
    (Corresponding negative statements including that without the O&G sector/company)
      - there would be a shortage or reduce amount of money/supplies going to charitable causes.

  CB: Focuses on the creation and sustainability of jobs by the oil and gas industry.
  # Here's the annotation criteria points:
    (General)
      - the O&G sector/company is providing jobs.
      - provides jobs paying high wages.
    (Corresponding negative statements including that without the O&G sector/company)
      - there would be job losses.
      - jobs would be lower paid.

Green Innovation and Climate Solutions
  GA: Highlights efforts to reduce greenhouse gas emissions through internal targets, policy support, voluntary initiatives, and emissions reduction technologies.
  # Here's the annotation criteria points:
    (The O&G sector/company is engaging in efforts to reduce greenhouse gas emissions)
      - setting internal targets.
      - supporting policy to reduce emissions.
      - supporting or participating in voluntary initiatives to reduce emissions.
      - supporting or developing means of emissions reduction including energy efficiency.
      - Carbon Capture, Utilisation and Storage (CCUS).
      - offsetting through planting trees.
    (Corresponding negative statements)
      - policy to reduce emissions is unnecessary or incorrect.

  GC: Promotes "clean" or "green" fossil fuels as part of climate solutions.
  # Here's the annotation criteria points:
    (General)
      - the O&G sector/company is developing 'clean' or 'green' fossil fuels (oil and gas).
      - oil and/or gas is 'clean' 'green' or 'sustainable'.
      - oil and/or gas is a partner the renewable energy.
      - oil and/or gas is a climate solution or part of the solution to climate change.

Pragmatism/Pragmatic Energy mix (Power systems and manufactured goods)
  PA: Portrays oil and gas as essential, reliable, affordable, and safe energy sources critical for maintaining power systems.
  # Here's the annotation criteria points:
    (General)
      - promoting the benefits of using oil and gas as energy sources.
      - claiming oil and gas are reliable, affordable, safe, efficient, etc.
      - suggesting that oil and gas are needed for the foreseeable future, that oil and gas are needed for or helps keeps the lights on, and that oil and gas powers people's lives (the implication being that it is needed for people's lives to continue as is or be powered effectively).
    (Corresponding negative statements)
      - without oil and gas energy would b unreliable, unaffordable, unsafe etc.

  PB: Emphasizes the importance of oil and gas as raw materials for various non-power-related uses and manufactured goods.
  # Here's the annotation criteria points:
    (General)
      - about alternative uses of oil and gas including to make PPE, toothbrushes etc.
    (Corresponding negative statements)
      - without oil and gas there wouldn't be or there would be a shortage in these goods.

Patriotic Energy mix
  SA: Stresses how domestic oil and gas production benefits the nation, including energy independence, energy leadership, and the idea of supporting American energy.
  # Here's the annotation criteria points:
    (General)
      - the O&G sector/company contributes to energy independence/national security.
    (Corresponding negative statements including that without the O&G sector/company)
      - the energy independence would suffer.
      - national security would suffer.
      - the US would be reliant on foreign/unstable/hostile regimes.

This task is a multi-label classification and can have up to the four labels.
Return a JSON list with all relevant labels. For example, a label containing "CA" and "PB" should be answered ["CA", "PB"]. If there is no label to annotate, return an empty list: [].

# Question and answering format

Frame & transcript (if available) pairs of the video: the video frames and transcripts.
Answer: the JSON list including annotated labels.

----------
```

Figure 12: The guideline part of the prompt for FACEBOOK.

Table 7: Entity list of YOUTUBE

Name	Country by headquarters	Region	# Video
Abu Dhabi National Oil Company	AE	Middle East	20
Alinta Energy	AU	Oceania	13
Ampol Limited	AU	Oceania	9
BP plc	GB	Europe	14
Bharat Petroleum Corporation Limited	IN	Asia	8
Cenovus Energy Inc.	CA	North America	1
Chevron Corporation	US	North America	14
China National Offshore Oil Corporation	CN	Asia	2
China Petroleum & Chemical Corporation (Sinopec)	CN	Asia	20
ConocoPhillips	US	North America	9
Coterra Energy Inc.	US	North America	8
ENEOS Corporation	JP	Asia	6
Ecopetrol S.A.	CO	South America	2
Enbridge Inc.	CA	North America	24
Equinor ASA	NO	Europe	4
Exxon Mobil Corporation	US	North America	12
Galp Energia SGPS, S.A.	PT	Europe	2
Halliburton Company	US	North America	5
INPEX Corporation	JP	Asia	2
Indian Oil Corporation	IN	Asia	6
Kinder Morgan, Inc.	US	North America	1
Marathon Petroleum Corporation	US	North America	3
Neste Oyj	FI	Europe	20
OMV Aktiengesellschaft	AT	Europe	19
Occidental Petroleum Corporation	US	North America	3
Oil and Natural Gas Corporation	IN	Asia	2
Origin Energy Limited	AU	Oceania	14
PKN Orlen	PL	Europe	3
PTT Public Company Limited	TH	Asia	6
Phillips 66 Company	US	North America	6
Pioneer Natural Resources Company	US	North America	1
Repsol S.A.	ES	Europe	21
S-OIL Corporation	KR	Asia	1
SLB (Schlumberger Limited)	US	North America	5
SNAM S.p.A.	IT	Europe	7
Saudi Aramco	SA	Middle East	23
Shell plc	GB	Europe	18
Suncor Energy Inc.	CA	North America	1
TC Energy Corporation	CA	North America	12
TotalEnergies SE	FR	Europe	18
Valero Energy Corporation	US	North America	16
Woodside Energy Group Ltd	AU	Oceania	5

A.12 Dataset Documentation

Based on the dataset sheets by Gebru et al. [19], the following provides information of our dataset.

A.12.1 Motivation

For what purpose was the dataset created? We created this annotated video dataset to benchmark VLMs for predicting obstruction and impressionistic framing by O&G entities.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)? A research team of the authors from Stanford University, Princeton University, InfluenceMap, and Hitachi America.

Who funded the creation of the dataset? No entity explicitly funded the creation of the dataset, but GM conducted this study within the Stanford Data Science (SDS) Affiliates Program.

A.12.2 Composition

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? The dataset includes videos, their metadata and annotated labels. We cover video publishers from different countries (at headquarters-level).

How many instances are there in total (of each type, if appropriate)? See Table 1 and Table 4.

```
As an expert annotator, please annotate "narratives" of advertising videos from the oil & gas (O&G) sector companies.

First, you must watch the video carefully (video frames and transcript (if available) will be given). Second, answer the labels based on the criteria if an audience will be impressed implicitly or explicitly by the video. You can select multiple labels. Video ad messages are aimed at a diverse range of people, and they usually convey a corporate impression without requiring deep thought. Therefore, when labeling, please try to consider implicit messages and follow your gut feeling after watching the video. Some labels may overlap, so please label all relevant labels.

# Label definitions

- "Economy and Business"
  - Impressions the audience will have: the company or O&G contributes to the local or national economy, industry, business, and tax revenue.
  - Example criteria to annotate: Narratives of tax income. Images of the number of investing. A testimony from local business owners. Images of industry serving a country. Images of the urban city, energy and transportation network, global connectivity map, and futuristic economy. Pictures of supply chains.
  - The following **should not** be considered: Projects, businesses, and investments that are not directly linked to O&G. General image of workers.

- "Work"
  - Impressions the audience will have: The company's employees are reliable. The workplace of the company is reliable. The company or O&G contributes to creating or sustaining jobs.
  - Example criteria to annotate: An interview with an employee. Images of Employees who work and employees who smile. Footage of plants, rigs and pipelines. Narratives of the number of jobs created.
  - The following **should not** be considered: Workers not linked to the O&G business.

- "Environment"
  - Impressions the audience will have: The company contributes to reducing GHG emissions. The company supports renewable energy. The company does good for the environment. The company uses "clean" or "sustainable" O&G.
  - Example criteria to annotate: Narratives like "We contribute to reducing CO2 emissions." Images of renewable energies like solar panels and wind turbines. Images of the low-emission vehicles or EV. Mountains, lakes, or any other beautiful natural images along with narratives that imply the company does good for the environment.
  - The following **should not** be considered: General images of beautiful landscape that do not associated with environmental impressions.

- "Green Innovation"
  - Impressions the audience will have: The company is innovative or futuristic to derive green future. The company is developing new energy technologies (which will be efficient.) A showcase of new renewable energy technology..
  - Example criteria to annotate: Images of research lab, developing new green solutions.
  - The following **should not** be considered: Technologies not linked to the O&G business.

- "Community and Life"
  - Impressions the audience will have: The company or O&G contributes to the community, culture, transportation, and sport. The company supports charitable efforts. The company or O&G supports our daily lives. The company or O&G supports one's hobby.
  - Example criteria to annotate: Images of local school and hospital. Narratives about sponsorship of a racing car. Narratives of local reforestation. Narratives for pragmatism like "O&G heats their homes". Images of a family smiling, eating, and playing.  Footage of daily life of people. Narratives of the need of oil or gas for cooking or daily plastic goods.

- "Patriotism"
  - Impressions the audience will have: The company or O&G contributes to the country. The company is a pride of our country. The production of O&G reserves benefits the country or energy independence.
  - Example criteria to annotate: Images of national flags. Emphasize national brands, e.g., "we produce the US brand."

This task is a multi-label classification and can have up to the six labels.
Return a JSON list with all relevant labels. For example, a label containing "Economy and Business" and "Community and Life" should be answered ["Economy and Business", "Community and Life"]. If there is no label to annotate, return an empty list: [].

# Question and answering format

Frame & transcript (if available) pairs of the video: the video frames and transcripts.
Answer: the JSON list including annotated labels.

------------------------
```

Figure 13: The guideline part of the prompt for YOUTUBE.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The video collection process includes sampling. In particular, videos were collected from the YOUTUBE domain by searching official channels, which involves a black box process. For details, see Appendix A.5.4.

What data does each instance consist of? Our dataset mainly consists of video metadata (such as URL) and annotated labels. See Appendix A.5.1 and Appendix A.5.3 for more detail.

Is there a label or target associated with each instance? Yes, the framing labels are associated with each video instance.

Is any information missing from individual instances? N/A.

Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? N/A.

Are there recommended data splits (for example, training, development/validation, testing)? Yes. There is an official data split for training and testing.

Are there any errors, sources of noise, or redundancies in the dataset? The annotations can contain errors by human mistakes or unrecognized biases. Some videos can be similar, e.g., an entity publishes similar videos based on a series content.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? The video content can be accessible by a link.

Table 8: Ablation results in the 1-shot prediction. T represents the transcript input, ES represents the embedding search in the 1-shot sample selection, and ER represents the entity restriction in the 1-shot sample selection.

Model	Params	T	ES	ER	YOUTUBE	FACEBOOK
DeepSeekVL2	4.5B	✓	✓	✓	49.7	62.3
		×	✓	✓	**53.5**	**63.8**
		✓	×	✓	46.9	43.8
		✓	✓	×	50.5	60.6
InternVL2	8B	✓	✓	✓	56.7	46.2
		×	✓	✓	**58.3**	**48.4**
		✓	×	✓	52.1	39.6
		✓	✓	×	56.3	46.1
Qwen2.5-VL	7B	✓	✓	✓	59.2	58.2
		×	✓	✓	**60.0**	53.6
		✓	×	✓	54.3	48.2
		✓	✓	×	57.3	**58.4**
Qwen2.5-VL	32B	✓	✓	✓	**66.2**	**70.5**
		×	✓	✓	61.2	60.6
		✓	×	✓	64.0	59.1
		✓	✓	×	65.6	68.1

Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset identify any subpopulations (for example, by age, gender)? No.

Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? No.

Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No.

A.12.3 Collection process

How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/ derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? The annotated labels for videos were originally annotated by expert annotators. Each video was directly observable to the annotators. For the FACEBOOK domain, it contains distant labels where we map the annotated labels on textual content to the video content automatically. See Section 2 for more detail.

What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? See Section 2.

If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)? To extract target videos for annotations of YOUTUBE, we sampled videos from the pool of available videos. The random sampling was applied for this process.

Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)? For YOUTUBE, the authors engaged in the annotations. There are no compensations.

Over what timeframe was the data collected? See Section 2 and Section 3.

Were any ethical review processes conducted (for example, by an institutional review board)? No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)? We obtained video data from websites.

Were the individuals in question notified about the data collection? N/A.

Did the individuals in question consent to the collection and use of their data? N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? N/A.

A.12.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? The labeling was conducted by annotators. We transcribe the videos using Whisper-1.

Was the "raw" data saved in addition to the preprocessed/cleaned/ labeled data (for example, to support unanticipated future uses)? Yes.

Is the software that was used to preprocess/clean/label the data available? No.

A.12.5 Uses

Has the dataset been used for any tasks already? No.

Is there a repository that links to any or all papers or systems that use the dataset? No.

What (other) tasks could the dataset be used for? The dataset might be used for potential task of corporate climate engagement assessment and greenwashing detection.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses? See Appendix A.2.

Are there tasks for which the dataset should not be used? This dataset should not be used to attack or mislabel real entities, locations, or individuals.

A.12.6 Distribution

Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? Yes.

How will the dataset be distributed (for example, tarball on website, API, GitHub)? See Appendix A.3.

When will the dataset be distributed? Before the camera ready deadline of this paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? Except for third-party content, we will license the dataset with CC BY-NC 4.0.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? Each video content is copyrighted by its respective publisher.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

A.12.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The authors.

How can the owner/curator/ manager of the dataset be contacted (for example, email address)? By the email address.

Is there an erratum? N/A.

Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? Yes. We may delete instances that associate with deleted videos.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A.

Will older versions of the dataset continue to be supported/hosted/ maintained? No.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? No.